

Best of LessWrong: September 2019

1. [Heads I Win, Tails?—Never Heard of Her; Or, Selective Reporting and the Tragedy of the Green Rationalists](#)
2. [The Zettelkasten Method](#)
3. [The unexpected difficulty of comparing AlphaStar to humans](#)
4. [Honoring Petrov Day on LessWrong, in 2019](#)
5. [Rationality Exercises Prize of September 2019 \(\\$1,000\)](#)
6. [Follow-Up to Petrov Day, 2019](#)
7. [What is operations?](#)
8. [Value Impact](#)
9. [System 2 as working-memory augmented System 1 reasoning](#)
10. [Bioinfohazards](#)
11. [Don't depend on others to ask for explanations](#)
12. [The Power to Judge Startup Ideas](#)
13. [A simple environment for showing mesa misalignment](#)
14. [AI Safety "Success Stories"](#)
15. [Tiddlywiki for organizing notes and research](#)
16. [Reframing the evolutionary benefit of sex](#)
17. [Free Money at PredictIt?](#)
18. [Reframing Impact](#)
19. [The Power to Demolish Bad Arguments](#)
20. [Utility ≠ Reward](#)
21. [The Inadequacy of Current Science \(Novum Organum Book 1: 1-37\)](#)
22. [Novum Organum: Introduction](#)
23. [Age gaps and Birth order: Reanalysis](#)
24. [How Specificity Works](#)
25. [\[Talk\] Paul Christiano on his alignment taxonomy](#)
26. [The Power to Be Emotionally Mature](#)
27. [Integrating the Lindy Effect](#)
28. [Three Stories for How AGI Comes Before FAI](#)
29. [Divergence on Evidence Due to Differing Priors - A Political Case Study](#)
30. [The Power to Teach Concepts Better](#)
31. [On Becoming Clueless](#)
32. [The strategy-stealing assumption](#)
33. [Candy for Nets](#)
34. [Gears vs Behavior](#)
35. [Long-term Donation Bunching?](#)
36. [How good is the case for retraining yourself to sleep on your back?](#)
37. [Specificity: Your Brain's Superpower](#)
38. [Idols of the Mind Pt. 1 \(Novum Organum Book 1: 38-52\)](#)
39. [Realism and Rationality](#)
40. [Seven habits towards highly effective minds](#)
41. [Focus](#)
42. [A Critique of Functional Decision Theory](#)
43. [What are the biggest "moonshots" currently in progress?](#)
44. [Timer Toxicities](#)
45. [Who's an unusual thinker that you recommend following?](#)
46. [Idols of the Mind Pt. 2 \(Novum Organum Book 1: 53-68\)](#)
47. [Concrete experiments in inner alignment](#)
48. [SSC Meetups Everywhere: Community Map on frontpage this month](#)
49. [Deducing Impact](#)
50. [Is Specificity a Mental Model?](#)

Best of LessWrong: September 2019

1. [Heads I Win, Tails?—Never Heard of Her; Or, Selective Reporting and the Tragedy of the Green Rationalists](#)
2. [The Zettelkasten Method](#)
3. [The unexpected difficulty of comparing AlphaStar to humans](#)
4. [Honoring Petrov Day on LessWrong, in 2019](#)
5. [Rationality Exercises Prize of September 2019 \(\\$1,000\)](#)
6. [Follow-Up to Petrov Day, 2019](#)
7. [What is operations?](#)
8. [Value Impact](#)
9. [System 2 as working-memory augmented System 1 reasoning](#)
10. [Bioinfohazards](#)
11. [Don't depend on others to ask for explanations](#)
12. [The Power to Judge Startup Ideas](#)
13. [A simple environment for showing mesa misalignment](#)
14. [AI Safety "Success Stories"](#)
15. [Tiddlywiki for organizing notes and research](#)
16. [Reframing the evolutionary benefit of sex](#)
17. [Free Money at PredictIt?](#)
18. [Reframing Impact](#)
19. [The Power to Demolish Bad Arguments](#)
20. [Utility ≠ Reward](#)
21. [The Inadequacy of Current Science \(Novum Organum Book 1: 1-37\)](#)
22. [Novum Organum: Introduction](#)
23. [Age gaps and Birth order: Reanalysis](#)
24. [How Specificity Works](#)
25. [\[Talk\] Paul Christiano on his alignment taxonomy](#)
26. [The Power to Be Emotionally Mature](#)
27. [Integrating the Lindy Effect](#)
28. [Three Stories for How AGI Comes Before FAI](#)
29. [Divergence on Evidence Due to Differing Priors - A Political Case Study](#)
30. [The Power to Teach Concepts Better](#)
31. [On Becoming Clueless](#)
32. [The strategy-stealing assumption](#)
33. [Candy for Nets](#)
34. [Gears vs Behavior](#)
35. [Long-term Donation Bunching?](#)
36. [How good is the case for retraining yourself to sleep on your back?](#)
37. [Specificity: Your Brain's Superpower](#)
38. [Idols of the Mind Pt. 1 \(Novum Organum Book 1: 38-52\)](#)
39. [Realism and Rationality](#)
40. [Seven habits towards highly effective minds](#)
41. [Focus](#)
42. [A Critique of Functional Decision Theory](#)
43. [What are the biggest "moonshots" currently in progress?](#)
44. [Timer Toxicities](#)
45. [Who's an unusual thinker that you recommend following?](#)
46. [Idols of the Mind Pt. 2 \(Novum Organum Book 1: 53-68\)](#)
47. [Concrete experiments in inner alignment](#)
48. [SSC Meetups Everywhere: Community Map on frontpage this month](#)

49. [Deducing Impact](#)

50. [Is Specificity a Mental Model?](#)

Heads I Win, Tails?—Never Heard of Her; Or, Selective Reporting and the Tragedy of the Green Rationalists

Followup to: [What Evidence Filtered Evidence?](#)

In "[What Evidence Filtered Evidence?](#)", we are asked to consider a scenario involving a coin that is *either* biased to land Heads 2/3rds of the time, *or* Tails 2/3rds of the time. Observing Heads is 1 bit of evidence for the coin being Heads-biased (because the Heads-biased coin lands Heads with probability 2/3, the Tails-biased coin does so with probability 1/3, the likelihood ratio of these is $\frac{2}{1/3} = 2$, and $\log_2 2 = 1$), and analogously and respectively for Tails.

If such a coin is flipped ten times by someone who [doesn't make literally false statements](#), who then reports that the 4th, 6th, and 9th flips came up Heads, then the update to our beliefs about the coin depends on what *algorithm* the not-lying^[1] reporter used to decide to report those flips in particular. If they always report the 4th, 6th, and 9th flips *independently* of the flip outcomes—if there's no [evidential entanglement](#) between the flip outcomes and the choice of which flips get reported—then reported flip-outcomes can be treated the same as flips you observed yourself: three Headses is $3 * 1 = 3$ bits of evidence in favor of the hypothesis that the coin is Heads-biased. (So if we were initially 50:50 on the question of which way the coin is biased, our posterior odds after collecting 3 bits of evidence for a Heads-biased coin would be $2^3 : 1 = 8:1$, or a probability of $8/(1 + 8) \approx 0.89$ that the coin is Heads-biased.)

On the other hand, if the reporter mentions only and exactly the flips that came out Heads, then we can *infer* that the other 7 flips came out Tails (if they didn't, the reporter would have mentioned them), giving us posterior odds of $2^3 : 2^7 = 1:16$, or a probability of around 0.06 that the coin is Heads-biased.

So far, so standard. (You *did* [read the Sequences](#), right??) What I'd like to emphasize about this scenario today, however, is that while a Bayesian reasoner who *knows* the non-lying reporter's algorithm of what flips to report will never be misled by the selective reporting of flips, a Bayesian with *mistaken* beliefs about the reporter's decision algorithm can be misled *quite badly*: compare the 0.89 and 0.06 probabilities we just derived given the *same* reported outcomes, but different assumptions about the reporting algorithm.

If the coin gets flipped a sufficiently large number of times, a reporter whom you *trust* to be impartial (but isn't), can *make you believe anything she wants without ever telling a single lie*, just with [appropriate selective reporting](#). Imagine a very biased coin that comes up Heads 99% of the time. If it gets flipped ten thousand times, 100 of those flips will be Tails (in expectation), giving a selective reporter plenty of examples to point to if she wants to convince you that the coin is extremely Tails-biased.

Toy models about biased coins are instructive for constructing examples with explicitly calculable probabilities, but the same *structure* applies to any real-world situation where you're receiving evidence from other agents, and you have uncertainty about what algorithm is being used to determine what reports get to you. Reality is like the coin's bias; evidence and arguments are like the outcome of a particular flip. *Wrong* theories [will still have some valid arguments and evidence supporting them](#) (as even a very Heads-biased coin will come up Tails sometimes), but theories that are [less wrong](#) will have *more*.

If selective reporting is mostly due to the idiosyncratic [bad intent](#) of rare malicious actors, then you might hope for safety in [\(the law of large\)](#) numbers: if Helga in particular is systematically more likely to report Headses than Tailses that she sees, then her flip reports will diverge from everyone else's, and you can take that into account when reading Helga's reports. On the other hand, if selective reporting is mostly due to systemic *structural* factors that result in *correlated* selective reporting even among well-intentioned people who are being honest as best they know how, [2] then you might have a more serious problem.

"[A Fable of Science and Politics](#)" depicts a fictional underground Society polarized between two partisan factions, the Blues and the Greens. "[T]here is a 'Blue' and a 'Green' position on almost every contemporary issue of political or cultural importance." If human brains consistently understood [the is/ought distinction](#), then political or cultural alignment with the Blue or Green agenda wouldn't distort people's beliefs about reality. Unfortunately ... humans. (I'm not even going to finish the sentence.)

Reality itself isn't on anyone's side, but any particular fact, argument, [sign, or portent](#) might just so happen to be more easily construed as "supporting" the Blues or the Greens. The Blues want stronger marriage laws; the Greens want no-fault divorce. An [evolutionary psychologist](#) investigating [effects of kin-recognition mechanisms on child abuse by stepparents](#) might aspire to scientific objectivity, but being objective and *staying* objective is *difficult* when you're embedded in an [intelligent social web](#) in which in your work is going to be predictably championed by Blues and reviled by Greens.

Let's make another toy model to try to understand the resulting distortions on the Undergrounders' collective epistemology. Suppose Reality is a coin—no, not a coin, a three-sided die, [3] with faces colored blue, green, and gray. One-third of the time it comes up blue (representing a fact that is more easily construed as supporting the Blue narrative), one-third of the time it comes up green (representing a fact that is more easily construed as supporting the Green narrative), and one-third of the time it comes up gray (representing a fact that not even the worst ideologues know how to spin as "supporting" their side).

Suppose each faction has social-punishment mechanisms enforcing consensus internally. [Without loss of generality](#), take the Greens (with the understanding that everything that follows goes just the same if you swap "Green" for "Blue" and *vice versa*). [4] People observe rolls of the die of Reality, and can freely choose what rolls to report—except a resident of a Green city who reports more than 1 blue roll for every 3 green rolls is assumed to be a secret Blue Bad Guy, and faces increasing social punishment as their ratio of reported green to blue rolls falls below 3:1. (Reporting gray rolls is always safe.)

The punishment is typically *informal*: there's no *official* censorship from Green-controlled local governments, just a visible incentive gradient made out of social-media pile-ons, denied promotions, lost friends and mating opportunities, increased risk of being involuntarily committed to psychiatric prison, [5] &c. Even people who privately agree with dissident speech might participate in punishing it, the better to evade punishment themselves.

This scenario presents a problem for people who live in Green cities who want to make *and share* accurate models of reality. It's impossible to report every die roll (the only 1:1 scale map of the territory, is the territory itself), but it seems clear that the most generally useful models—the ones you would expect arbitrary AIs to come up with—aren't going to be sensitive to which facts are "blue" or "green". The reports of aspiring epistemic rationalists who are *just trying to make sense of the world* will end up being about one-third blue, one-third green, and one-third gray, matching the distribution of the Reality die.

From the perspective of ordinary nice smart Green citizens who have not been trained in [the Way](#), these reports look *unthinkably* Blue. Aspiring epistemic rationalists who are actually [paying attention](#) can easily distinguish Blue partisans from actual truthseekers, [6] but the [social-punishment machinery](#) can't process more than [five words at a time](#). The social consequences of being an *actual* Blue Bad Guy, or just an honest nerd who doesn't know when to keep her stupid trap shut, are the same.

In this scenario, [7] public opinion within a subculture or community in a Green area is constrained by the 3:1 (green:blue) "[Overton](#) ratio." In particular, under these conditions, it's *impossible to have a rationalist community*—at least the most naïve conception of such. If your marketing literature says, "Speak the truth, even if your voice trembles," but all the [savvy](#) high-status people's actual *reporting algorithm* is, "Speak the truth, except when that would cause the local social-punishment machinery to mark me as a Blue Bad Guy and hurt me and any people or institutions I'm associated with—in which case, tell the most convenient lie-of-omission", then [smart sincere](#) idealists who have internalized your marketing literature as a moral ideal and trust the community to implement that ideal, are going to be *misled* by the community's stated beliefs—and *confused* at some of the pushback they get when submitting reports with a 1:1:1 blue:green:gray ratio.

Well, misled to *some* extent—maybe not much! In the absence of an [Oracle AI](#) (or a competing rationalist community in Blue territory) to compare notes with, then it's not clear how one could get a *better* map than trusting what the "green rationalists" say. With a few more made-up modeling assumptions, we can *quantify* the distortion introduced by the Overton-ratio constraint, which will hopefully help develop an *intuition* for how large of a problem this sort of thing might be in real life.

Imagine that Society needs to make a decision about an Issue (like a question about divorce law or merchant taxes). Suppose that the facts relevant to making optimal decisions about an Issue are represented by nine rolls of the Reality die, and that the quality (utility) of Society's decision is proportional to the (base-two logarithm) entropy of the distribution of what facts get heard and discussed. [8]

The maximum achievable decision quality is $\log_2 9 \approx 3.17$.

On average, Green partisans will find 3 "green" facts [9] and 3 "gray" facts to report, and mercilessly [stonewall](#) anyone who tries to report any "blue" facts, for a decision

quality of $\log_2 6 \approx 2.58$.

On average, the Overton-constrained rationalists will report the same 3 "green" and 3 "gray" facts, but something interesting happens with "blue" facts: each individual can only afford to report one "blue" fact without blowing their Overton budget—but it doesn't have to be the *same* fact for each person. Reports of all 3 (on average) blue rolls get to enter the public discussion, but get mentioned (cited, retweeted, &c.) 1/3 as often as green or gray rolls, in accordance with the Overton ratio. So it turns out that the constrained rationalists end up with a decision quality of $\frac{1}{3} \log_2 7 + \frac{2}{3} \log_2 21 \approx 3.03$, [10] significantly better than the Green partisans—but still falling short of the theoretical ideal where all the relevant facts get their due attention.

If it's just not *pragmatic* to expect people to defy [their incentives](#), is this the best we can do? Accept a somewhat distorted state of discourse, forever?

At least one *partial* remedy seems apparent. Recall from our original coin-flipping example that a Bayesian who *knows* what the filtering process looks like, can take it into account and make the correct update. If you're filtering your evidence to avoid social punishment, but it's possible to clue in your fellow rationalists to your *filtering algorithm* without triggering the social-punishment machinery—you mustn't [assume that everyone already knows!](#)—that's potentially a big win. In other words, [blatant cherry-picking is the best kind!](#)

1. I don't quite want to use the word *honest* here. ↵
2. And it turns out that knowing *how* to be honest is *much more work* than one might initially think. You [have read the Sequences](#), right?! ↵
3. For lack of an appropriate [Platonic solid](#) in three-dimensional space, maybe imagine tossing a triangle in two-dimensional space?? ↵
4. As an author, I'm facing some conflicting desiderata in my color choices here. I want to say "Blues and Greens" *in that order* for consistency with "A Fable of Science and Politics" (and other [classics from the Sequences](#)). Then when making an arbitrary choice to talk in terms of one of the factions in order to avoid cluttering the exposition, you might have expected me to say "Without loss of generality, take the Blues," because the *first* item in a sequence ("Blues" in "Blues and Greens") is a more of a [Schelling point](#) than the second, or last, item. But I don't *want* to take the Blues, because that color choice [has other associations](#) that I'm trying to avoid right now: if I said "take the Blues", I fear many readers would assume that I'm trying to directly push a partisan point about [soft censorship](#) and [preference-falsification](#) social pressures in liberal/left-leaning subcultures in the contemporary United States. To be fair, it's *true* that soft censorship and preference-falsification social pressures in liberal/left-leaning subcultures in the contemporary United States are, historically, what inspired me, personally, to write this post. It's okay for you to notice that! But I'm *trying* to talk about the *general mechanisms* that generate this *class* of distortions on a Society's collective epistemology, independently of which faction or which ideology happens to be "on top" [in a particular place and time](#). If I'm *doing my job right*, then my analogue in a ["nearby" Everett branch](#) whose local subculture

was as "right-polarized" as my Berkeley environment is "left-polarized", would have written a post making the *same* arguments. ↵

5. Okay, they market themselves as psychiatric "hospitals", but let's not be confused by [misleading labels](#). ↵
6. Or rather, aspiring epistemic rationalists can do a *decent* job of assessing the *extent to which* someone is exhibiting truth-tracking behavior, or Blue-partisan behavior. Obviously, people who are *consciously* trying to seek truth, are not necessarily going to *succeed at* [overcoming bias](#), and attempts to correct for the "pro-Green" distortionary forces being discussed in this parable could easily veer into "pro-Blue" over-correction. ↵
7. Please be appropriately skeptical about the real-world relevance of my made-up modeling assumptions! If it turned out that my choice of assumptions were (subconsciously) selected for the resulting conclusions about how bad evidence-filtering is, that would be really bad for the same reason that I'm claiming that evidence-filtering is really bad! ↵
8. The entropy of a discrete probability distribution is maximized by the uniform distribution, in which all outcomes receive equal probability-mass. I only chose these "exactly nine equally-relevant facts/rolls" and "entropic utility" assumptions to make the arithmetic easy on me; a more realistic model might admit arbitrarily many facts into discussion of the Issue, but posit a distribution of facts/rolls with [diminishing marginal](#) relevance to Society's decision quality. ↵
9. The scare quotes around the adjective "'green'" (&c.) when applied to the word "fact" (as opposed to a die roll outcome *representing* a fact in our toy model) are significant! The facts aren't actually on anyone's side! We're trying to model the *distortions* that arise from stupid humans *thinking* that the facts are on someone's side! This is sufficiently important—and difficult to remember—that I should probably repeat it until it becomes obnoxious! ↵
10. You have three green slots, three gray slots, and three blue slots. You put three counters each on each of the green and gray slots, and one counter each on each of the blue slots. The frequencies of counters per slot is [3, 3, 3, 3, 3, 3, 1, 1, 1]. The total number of counters you put down is $3 \cdot 6 + 3 = 18 + 3 = 21$. To turn the frequencies into a probability distribution, you divide everything by 21, to get [1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/21, 1/21, 1/21]. Then the entropy is $6 \cdot -\frac{1}{7} \log_2 \frac{1}{7} + 3 \cdot -\frac{1}{7} \log_2 \frac{1}{7}$, which simplifies to $\frac{6}{7} \log_2 7 + \frac{3}{7} \log_2 21$. ↵

The Zettelkasten Method

[Epistemic Status: Scroll to the bottom for my follow-up thoughts on this from months/years later.]

Early this year, Conor White-Sullivan introduced me to the Zettelkasten method of note-taking. I would say that this significantly increased my research productivity. I've been saying "at least 2x". Naturally, this sort of thing is difficult to quantify. The truth is, I think it may be more like 3x, especially along the dimension of "producing ideas" and also "early-stage development of ideas". (What I mean by this will become clearer as I describe how I think about research productivity more generally.) However, it is also very possible that the method produces serious *biases* in the types of ideas produced/developed, which should be considered. (This would be difficult to quantify at the best of times, but also, it should be noted that other factors have dramatically *decreased* my overall research productivity. So, unfortunately, someone looking in from outside would not see an overall boost. Still, *my* impression is that it's been very useful.)

I think there are some specific reasons why Zettelkasten has worked so well for me. I'll try to make those clear, to help readers decide whether it would work for them. However, I honestly didn't think Zettelkasten sounded like a good idea before I tried it. It only took me about 30 minutes of working with the cards to decide that it was really good. So, if you're like me, this is a cheap experiment. I think a lot of people should actually try it to see how they like it, even if it sounds terrible.

My plan for this document is to first give a short summary and then an overview of Zettelkasten, so that readers know roughly what I'm talking about, and can possibly experiment with it without reading any further. I'll then launch into a longer discussion of why it worked well for me, explaining the specific habits which I think contributed, including some descriptions of my previous approaches to keeping research notes. I expect some of this may be useful even if you don't use Zettelkasten -- if Zettelkasten isn't for you, maybe these ideas will nonetheless help you to think about optimizing your notes. However, I put it here primarily because I think it will boost the chances of Zettelkasten working for you. It will give you a more concrete picture of how I use Zettelkasten as a thinking tool.

Very Short Summary

Materials

- [Staples index-cards-on-a-ring](#) or equivalent, possibly with:
 - [plastic rings](#) rather than metal
 - different 3x5 index cards (I recommend blank, but, other patterns may be good for you) as desired
 - some kind of divider
 - I use yellow index cards as dividers, but slightly larger cards, tabbed cards, plastic dividers, etc. might be better

- quality hole punch (if you're using different cards than the pre-punched ones)
 - I like [this one](#).
- Blank stickers or some other way to label card-binders with the address range stored within.
- quality writing instrument -- must suit you, but,
 - multi-color click pen recommended
 - hi-tec-c coleto especially recommended

Technique

- Number pages with alphanumeric strings, so that pages can be sorted hierarchically rather than linearly -- 11a goes between 11 and 12, 11a1 goes between 11a and 11b, *et cetera*. This allows pages to be easily inserted between other pages without messing up the existing ordering, which makes it much easier to continue topics.
- Use the alphanumeric page identifiers to “hyperlink” pages. This allows sub-topics and tangents to be easily split off into new pages, and also allows for related ideas to be interlinked.

Before I launch into the proper description of Zettelkasten, here are some other resources on note-taking which I looked at before diving into using Zettelkasten myself. (Feel free to skip this part on a first reading.)

Related Literature

There are other descriptions of Zettelkasten out there. I mainly read *How to Take Smart Notes*, which is the best book on Zettelkasten as far as I know -- it claims to be the best write-up available *in English*, anyway. The book contains a thorough description of the technique, plus a lot of “philosophical” stuff which is intended to help you approach it with the right mindset to actually integrate it into your thinking in a useful way. I am sympathetic to this approach, but some of the content seems like bad science to me (such as the description of growth mindset, which didn’t strike me as at all accurate -- I’ve read some of the original research on growth mindset).

An issue with some other write-ups is that they focus on implementing Zettelkasten-like systems digitally. In fact, Conor White-Sullivan, who I’ve already mentioned, is working on a Workflowy/Dynalist-like digital tool for thinking, inspired partially by Zettelkasten (and also by the idea that a Workflowy/Dynalist style tool which is *designed explicitly to nudge users into good thinking patterns* with awareness of cognitive biases, good practices for argument mapping, etc. could be very valuable). You can take a look at his tool, Roam, [here](#). He also wrote up some thoughts about [Zettelkasten in Roam](#). However, ***I strongly recommend trying out Zettelkasten on actual note-cards***, even if you end up implementing it on a computer. There’s something good about the note-card version that I don’t fully understand. As such, I would advise against trusting other people’s attempts to distill what makes Zettelkasten good into a digital format -- better to try it yourself, so that you can then judge whether alternate versions are improvements for you. The version I will describe here is fairly close to the original.

I don't strongly recommend my own write-up over what's said in *How to Take Smart Notes*, particularly the parts which describe the actual technique. I'm writing this up partly just so that there's an easily linkable document for people to read, and partly because I have some ideas about how to make Zettelkasten work for you (based on my own previous note-taking systems) which are different from the book.

Another source on note-taking which I recommend highly is Lion Kimbro's *How to Make a Complete Map of Every Thought You Think* ([html](#), [pdf](#)). This is about a completely different system of note-taking, with different goals. However, it contains a wealth of inspiring ideas about note-taking systems, including valuable tips for the raw physical aspects of keeping paper notes. I recommend reading [this interview with Lion Kimbro](#) as a "teaser" for the book -- he mentions some things which he didn't in the actual book, and it serves somewhat as "the missing introduction" to the book. (You can skip the part at the end about wikis if you don't find it interesting; it is sort of outdated speculation about the future of the web, and it doesn't get back to talking about the book.) Part of what I love about *How to Make a Complete Map of Every Thought You Think* is the manic brain-dump writing style -- it is a book which feels very "alive" to me. If you find its style grating rather than engaging, it's probably not worth you reading through.

I should also mention [another recent post about Zettelkasten here on LW](#).

Zettelkasten, Part 1: The Basics

Zettelkasten is German for ‘slip-box’, IE, a box with slips of paper in it. You keep everything on a bunch of note cards. Niklas Luhmann developed the system to take notes on his reading. He went on to be an incredibly prolific social scientist. It is hard to know whether his productivity was tied to Zettelkasten, but he thinks so, and others have reported large productivity boosts from the technique as well.

Small Pieces of Paper Are Just Modular Large Pieces of Paper

You may be thinking: aren’t small pieces of paper bad? Aren’t [large notebooks just better](#)? Won’t small pages make for small ideas?

What I find is that the drive for larger paper is better-served by splitting things off into new note cards. Note-cards relevant to your current thinking can be spread on a table to get the same big-picture overview which you’d get from a large sheet of paper. Writing on an *actual* large sheet of paper locks things into place.

When I was learning to write in my teens, it seemed to me that paper was a prison. Four walls, right? And the ideas were constantly trying to escape. What is a parenthesis but an idea trying to escape? What is a footnote but an idea that tried -- that jumped off the cliff? Because paper enforces single sequence -- and there’s no room for digression -- it imposes a particular kind of order in the very nature of the structure.

-- Ted Nelson, [demonstration of Xanadu space](#)

I use 3x5 index cards. That’s quite small compared to most notebooks. It may be that this is the right size for me only because I already have very small handwriting. I believe Luhmann used larger cards. However, *I expected it to be too small*. Instead, I found the small cards to be freeing. I strongly recommend trying 3x5 cards before trying with a larger size. In fact, even smaller sizes than this are viable -- one early reader of this write-up decided to use *half* 3x5 cards, so that they’d fit in mtg deck boxes.

Writing on small cards *forces* certain habits which would be good even for larger paper, but which I didn’t consider until the small cards made them necessary. It forces ideas to be [broken up into simple pieces](#), which helps to clarify them. Breaking up ideas forces you to link them together explicitly, rather than relying on the linear structure of a notebook to link together chains of thought.

Once you’re forced to adopt a linking system, it becomes natural to use it to “break out of the prison of the page” -- tangents, parentheticals, explanatory remarks, caveats, ... everything becomes a new card. This gives your thoughts much more “surface area” to expand upon.

On a computer, this is essentially the wiki-style [[magic link]] which links to a page if the page exists, or creates the page if it doesn't yet exist -- a [critical but all-too-rare](#) feature of note-taking software. Again, though, I strongly recommend trying the system on paper before jumping to a computer; putting yourself in a position where you *need* to link information like crazy will help you to see the value of it.

This brings us to one of the defining features of the Zettelkasten method: the addressing system, which is how links between cards are established.

Paper Hypertext

We want to use card addresses to organize and reference everything. So, when you start a new card, its address should be the first thing you write -- you never want to have a card go without an address. Choose a consistent location for the addresses, such as the upper right corner. If you're using multi-color pens, like me, you might want to choose one color just for addresses.

Wiki-style links tend to use the title of a page to reference that page, which works very well on a computer. However, for a pen-and-paper hypertext system, we want to optimize several things:

- Easy lookup: we want to find referenced cards as easily as possible. This entails sorting the cards, so that you don't have to go digging; finding what you want is as easy as finding a word in the dictionary, or finding a page given the page number.
- Easy to sort: I don't know about you, but for me, putting things in alphabetical order isn't the easiest thing. I find myself reciting the alphabet pretty often. So, I don't *really* want to sort cards alphabetically by title.
- Easy to write: another reason not to sort alphabetically by title is that you want to reference cards really easily. You probably don't want to write out full titles, unless you can keep the titles really short.
- Fixed addresses: Whatever we use to reference a card, it must remain fixed. Otherwise, references could break when things change. No one likes broken links!
- [Related cards should be near each other](#). Alphabetical order might put closely related cards very far apart, which gets to be cumbersome as the collection of cards grows -- even if look-up is quite convenient, it is nicer if the related cards are already at hand without purposefully deciding to look them up.
- [No preset categories](#). Creating a system of categories is a common way to place related content together, but, it is too hard to know how you will want to categorize everything ahead of time, and the needs of an addressing system make it too difficult to change your category system later.

One simple solution is to number the cards, and keep them in numerical order. Numbers are easy to sort and find, and are very compact, so that you don't have the issue of writing out long names. However, although related content will be somewhat nearby (due to the fact that we're likely to create several cards on a topic at the same time), we can do better.

The essence of the Zettelkasten approach is the use of repeated decimal points, as in "22.3.14" -- cards addressed 2.1, 2.2, 2.2.1 and so on are all thought of as

“underneath” the card numbered 2, just as in the familiar subsection-numbering system found in many books and papers. This allows us to insert cards anywhere we want, rather than only at the end, which allows related ideas to be placed near each other much more easily. A card sitting “underneath” another can loosely be thought of as a comment, or a continuation, or an associated thought.

However, for the sake of compactness, Zettelkasten addresses are usually written in an alphanumeric format, so that rather than writing 1.1.1, we would write 1a1; rather than writing 1.2.3, we write 1b3; and so on. This notation allows us to avoid writing so many periods, which grows tiresome.

Alternating between numbers and letters in this way allows us to get to two-digit numbers (and even two-digit letters, if we exhaust the whole alphabet) without needing periods or dashes or any such separators to indicate where one number ends and the next begins.

Let’s say I’m writing linearly -- something which could go in a notebook. I might start with card 11, say. Then I proceed to card 11a, 11b, 11c, 11d, etc. On each card, I make a note somewhere about the previous and next cards in sequence, so that later I know for sure how to follow the chain via addresses.

Later, I might have a different branch-off thought from 11c. This becomes 11c1. That’s the magic of the system, which you can’t accomplish so easily in a linear notebook: you can just come back and add things. These tangents can grow to be larger than the original.



Don't get too caught up in what address to give a card to put it near relevant material. A card can be put anywhere in the address system. The point is to make things more convenient for you; nothing else matters. Ideally, the tree would perfectly reflect some kind of conceptual hierarchy; but in practice, card 11c might turn out to be the primary thing, with card 11 just serving as a historical record of what seeded the idea.

Similarly, a linear chain of writing doesn't have to get a nice linear chain of addresses. I might have a train of thought which goes across cards 11, 11a, 11b, 11b1, 11b1a, 11b1a1, 18, 18a... (I write a lot of "1alalala", and it is sometimes better to jump up to a new top-level number to keep the addresses from getting longer.)

Mostly, though, I've written less and less in linear chains, and more and more in branching trees. Sometimes a thought just naturally wants to come out linearly. But, this tends to make it more difficult to review later -- the cards aren't split up into atomic ideas, instead flowing into each other.

If you don't know where to put something, make it a new top-level card. You can link it to whatever you need via the addressing system, so the cost of putting it in a suboptimal location isn't worth worrying about too much! You don't want to be constrained by the ideas you've had so far. Or, to put it a different way: it's like starting a new page in a notebook. Zettelkasten is supposed to be *less* restrictive than a notebook, not more. Don't get locked into place by trying to make the addresses perfectly reflect the logical organization.

Physical Issues: Card Storage

Linear notes can be kept in any kind of paper notebook. Nonlinear/modular systems such as Zettelkasten, on the other hand, require some sort of binder-like system where you can insert pages at will. I've tried a lot of different things. Binders are typically just *less comfortable* to write in (because of the rings -- this is another point where the fact that I'm left-handed is very significant, and right-handed readers may have a different experience).

(One thing that's improved my life is realizing that I can use a binder "backwards" to get essentially the right-hander's experience -- I write on the "back" of pages, starting from the "end".)

They're also bulky; it seems somewhat absurd how much more bulky they are than a notebook of equivalently-sized paper. This is a serious concern if you want to carry them around. (As a general rule, I've found that a binder feels roughly equivalent to one-size-larger notebook -- a three-ring binder for 3x5 cards feels like carrying around a deck of 4x6 cards; a binder of A6 paper feels like a notebook of A5 paper; and so on.)

Index cards are often kept in special boxes, which you can get. However, I don't like this so much? I want a more binder-like thing which I can easily hold in my hands and flip through. Also, boxes are often made to view cards in landscape orientation, but I prefer portrait orientation -- so it's hard to flip through things and read while they're still in the box.

Currently, I use the [Staples index-cards-on-a-ring](#) which put all the cards on a single ring, and protect them with plastic covers. However, I replace the metal rings (which I find harder to work with) with [plastic rings](#). I also bought a variety of note cards to try -- you can try thicker/thinner paper, colors, line grid, dot grid, etc. If you do this, you'll need a hole punch, too. I recommend getting a "low force" hole punch; if you just go and buy the cheapest hole punch you can find, it'll probably be pretty terrible. You want to be fairly consistent with where you punch the holes, but, that wasn't as important as I expected (it doesn't matter as much with a one-ring binder in contrast to a three-ring, since you're not trying to get holes to line up with each other).

I enjoy the ring storage method, because it makes cards really easy to flip through, and I can work on several cards at once by splaying them out (which means I don't lose my place when I decide to make a new card or make a note on a different one, and don't have to take things out of sort order to work with them).

Deck Architecture

I don't keep the cards perfectly sorted all the time. Instead, I divide things up into sorted and not-yet-sorted:



(Blue in this image mean “written on” -- they’re all actually white except for the yellow divider, although of course you could use colored cards if you like.)

Fetch Modi

As I write on blank cards, I just leave them where they are, rather than immediately putting them into the sort ordering. I sort them in later. (Unsorted cards still have addresses and can be referenced. *The address is always, always the very first thing I write on a card.*)

There is an advantage to this approach beyond the efficiency of sorting things all at once. The unsorted cards are a physical record of *what I’m actively working on*. Since cards are so small, working on an idea almost always means creating new cards. So, I can easily jump back into whatever I was thinking about last time I handled the binder of cards.

Unless you have a specific new idea you want to think about (in which case you start a new card, or, go find the most closely related cards in your existing pile), there are basically two ways to enter into your card deck: from the front, and from the back. The front is “top-down” (both literally and figuratively), going from bigger ideas to smaller details. It’s more breadth-first. You’re likely to notice an idea which you’ve been neglecting, and start a new branch from it. Starting from the back, on the other hand, is depth-first. You’re continuing to go deeper into a branch which you’ve already developed some depth in.

Don’t sort too often. The unsorted cards are a valuable record of what you’ve been thinking about. I’ve regretted sorting too frequently -- it feels like I have to start over, find the interesting open questions buried in my stack of cards all over again.

In theory, one could also move cards from sorted to unsorted specifically to remind oneself to work on those cards, but I haven’t really used this tactic.

The advantage of sorting is to make address lookup easier. But, actually, address lookup in my unsorted cards is not that hard! Because the cards remain in creation-order, I know that e.g. card 10a1 must come *somewhere* after card 10a. It just doesn’t need to be *immediately* after when the cards aren’t sorted.

Splitting & Deck Management

When a ring feels over-full (after I fill approximately 100 cards), I sort all of the cards, and split the deck into two. (Look for a sensible place to split the tree into two -- you want to avoid a deep branch being split up into two separate decks, as much as you can.) Load up the two new decks with 50ish blank cards each, and stick them on new rings.

Everything is still on one big addressing system, so, it is a good idea to label the two new binders with the address range within. I use blank stickers, which I put on the front of each ring binder. The labels serve both to keep lookup easy (I don’t want to be guessing about which binder certain addresses are in), and also, to remind me to limit the addresses within a given deck.

For example, suppose this is my first deck of cards (so before the split, it holds everything). Let's say there are 30 cards underneath "1", 20 cards underneath "2", and then about 50 more cards total, under the numbers 3 through 14.

I would split this deck into a "1 through 2" deck, and a "3 through *" deck -- the * meaning "anything". You might think it would be "3 through 14", but, when I make card 15, it would go in that deck. So at any time, you have one deck of cards with no upper bound. On the other hand, when you are working with the "1 - 2" deck, you don't want to mistakenly make a card 3; you've already got a card 3 somewhere. You don't want duplicate addresses anywhere!

Currently, I have 6 decks: 0 - 1.4, 1.5 - 1.*, 2 - 2.4, 2.5 - 2.*, 3, and 4 - 4.*. (I was foolish when I started my Zettelkasten, and used the decimal system rather than the alphanumeric system. I switched quickly, but all my top-level addresses are still decimal. So, I have a lot of mixed-address cards, such as 1.3a1, 1.5.2a2, 2.6b4a, etc. As for why my numbers start at 0 rather than 1, I'll discuss that in the "Index & Bibliography" section.)

I like to have the unsorted/blank "short-term memory" section on every single deck, so that I can conveniently start thinking about stuff within that deck without grabbing anything else. However, it might also make sense to have only one "short-term memory" in order to keep yourself more focused (and so that there's only one place to check when you want to remember what you were recently working on!).

Getting Started: Your First Card

[Your first note](#) doesn't need to be anything important -- it isn't as if every idea you put into your Zettelkasten has to be "underneath" it. Remember, you aren't trying to invent [a good category system](#). Not every card has to look like a core idea with bullet points which elaborate on that idea, like my example in the previous section. You can just start writing whatever. In fact, it might be good if you make your first cards messy and unimportant, just to make sure you don't feel like everything has to be nicely organized and highly significant.

On the other hand, it might be important to have a good starting point, if you really want to give Zettelkasten a chance.

I mentioned that I knew I liked Zettelkasten within the first 30 minutes. I think it might be important that when I sat down to try it, I had an idea I was excited to work on. It wasn't a nice solid mathematical idea -- it was a fuzzy idea, one which had been burning in the back of my brain for a week or so, waiting to be born. It filled the fractal branches of a zettelkasten nicely, expanding in every direction.

So, maybe start with one of *those* ideas. Something you've been struggling to articulate. Something which hasn't found a place in your linear notebook.

Alright. That's all I have to say about the basics of Zettelkasten. You can go try it now if you want, or keep reading. The rest of this document is about further ideas in note-taking which have shaped the way I use Zettelkasten. These may or may not be useful to you; I don't know for sure why Zettelkasten is such a productive system for me personally.

Note-Taking Systems I Have Known and Loved

I'm organizing this section by my previous note-taking systems, but secretly, the main point is to convey a number of note-taking ideas which may have contributed to Zettelkasten working well for me. These ideas have seemed generally useful to me -- maybe they'll be useful to you, even if you don't end up using Zettelkasten in particular.

Notebooks

Developing Ideas

Firstly, and most importantly, I have been keeping idea books since middle school. I think there's something very important in the simple idea of writing regularly -- I don't have the reference, but, I remember reading someone who described the day they first started keeping a diary as the day they first woke up, started reflectively thinking about their relationship with the world. Here's a somewhat similar quote from a Zettelkasten blog:

During the time spanning Nov. 2007-Jan. 2010, I filled 11 note books with ideas, to-do lists, ramblings, diary entries, drawings, and worries.

Looking back, this is about the time I started to live consciously. I guess keeping a journal helped me "wake up" from some kind of teenage slumber.

--Christian

I never got into autobiographical diary-style writing, personally, instead writing about ideas I was having. Still, things were in a very "narrative" format -- the ideas were a drama, a back-and-forth, a dance of rejoinders. There was some math -- pages filled with equations -- but only after a great deal of (very) informal development of an idea.

As a result, "elaborate on an idea" / "keep going" seems like a primitive operation to me -- and, specifically, a primitive operation *which involves paper*. (I can't translate the same thinking style to conversation, not completely.) I'm sure that there is a lot to unpack, but for me, it just feels natural to keep developing ideas further.

So, when I say that the Zettelkasten card 1b2 "*elaborates on*" the card 1b, I'm calling on the long experience I've had with idea books. I don't know if it'll mean the same thing for you.

Here's my incomplete attempt to convey some of what it means.

When I'm writing in an idea book, I spend a lot of time trying to clearly explain ideas under the (often false) assumption that I know what I'm talking about. There's an imaginary audience who knows a lot of what I'm talking about, but I have to explain

certain things. I can't get away with leaving important terms undefined -- I have to establish anything I feel less than fully confident about. For example, the definition of a Bayesian network is something I can assume my "audience" can look up on wikipedia. However, if I'm less than totally confident in the concept of d-separation, I have to explain it; especially if it is important to the argument I hope to make.

Once I've established the terms, I try to explain the idea I was having. I spend a lot of time staring off into space, not really knowing what's going on in my head exactly, but with a sense that there's a simple point I'm trying to make, if only I could see it. I simultaneously feel like I know what I want to say (if only I could find the words), and like I don't know what it is -- after all, I haven't articulated it yet. Generally, I can pick up where I left off with a particular thought, even after several weeks -- I can glance at what I've written so far, and get right back to staring at the wall again, trying to articulate the same un-articulated idea.

If I start again in a different notebook (for example, switching to writing my thoughts on a computer), I have to explain everything again. This audience doesn't know yet! I can't just pick up on a computer where I left off on paper. It's like trying to pick up a conversation in the middle, but with a different person. This is sort of annoying, but often good (because re-explaining things may hold surprises, as I notice new details.)

Similarly, if I do a lot of thinking without a notebook (maybe in a conversation), I generally have to "construct" my new position from my old one. This has an unfortunate "freezing" effect on thoughts: there's a lot of gravity toward the chain of thought wherever it is on the page. I tend to work on whatever line of thought is most recent in my notebook, regardless of any more important or better ideas which have come along -- especially if the line of thought in the notebook isn't yet at a conclusive place. Sometimes I put a scribble in the notebook after a line of thought, to indicate explicitly that it no longer reflects the state of my thinking, to give myself "permission" to do something else.

Once I've articulated some point, then criticisms of the point often become clear, and I'll start writing about them. I often have a sense that I know how it's going to go a few steps ahead in this back-and-forth; a few critiques and replies/revisions. Especially if the ideas are flowing faster than I can write them down. However, it is important to actually write things down, because they often don't go quite as I expect.

If an idea seems to have reached a natural conclusion, including all the critiques/replies which felt important enough to write, I'll often write a list of "future work": any open questions I can think of, applications, details which are important but not so important that I want to write about them yet, etc. At this point, it is usually time to write the idea up for a *real* audience, which will require more detail and refine the idea yet further (possibly destroying it, or changing it significantly, as I often find a critical flaw when I try to write an idea up for consumption by others).

If I don't have any particular idea I'm developing, I may start fresh with a mental motion like "OK, obviously I know how to solve everything" and write down the grand solution to everything, starting big-picture and continuing until I get stuck. Or, instead, I might make a bulleted list free-associating about what I think the interesting problems are -- the things I don't know how to do.

Workflowy

The next advance in my idea notes was workflowy. I still love the simplicity of workflowy, even though I have moved on from it.

For those unfamiliar, Workflowy is an outlining tool. I was unfamiliar with the idea before Workflowy introduced it to me. Word processors generally *support* nested bulleted lists, but the page-like format of a word processor limits the depth such lists can go, and it didn't really occur to me to use these as a primary mode of writing. Workflowy doesn't let you do anything *but* this, and it provides enough features to make it extremely convenient and natural.

Nonlinear Ideas: Branching Development

Workflowy introduced me to the possibility of nonlinear formats for idea development. I've already discussed this to some extent, since it is also one of the main advantages of Zettelkasten over ordinary notebooks.

Suddenly, I could continue a thread *anywhere*, rather than always picking it up at the end. I could sketch out where I expected things to go, with an outline, rather than keeping all the points I wanted to hit in my head as I wrote. If I got stuck on something, I could write about how I was stuck nested *underneath* whatever paragraph I was currently writing, but then collapse the meta-thoughts to be invisible later -- so the overall narrative doesn't feel interrupted.

In contrast, writing in paper notebooks forces you to choose consciously that you're done for now with a topic if you want to start a new one. Every new paragraph is like choosing a single fork in a twisting maze. Workflowy allowed me to take them all.

What are Children?

I've seen people hit a block right away when they try to use workflowy, because they don't know what a "child node" is.

- Here's a node. It could be a paragraph, expressing some thought. It could also be a title.
 - Here's a child node. It could be a comment on the thought -- an aside, a critique, whatever. It could be something which goes under the heading.
- Here's a sibling node. It could be the next paragraph in the "main thrust" of an argument. It could be an unrelated point under the same super-point everything is under.

As with Zettelkasten, my advice is to not get too hung up on this. A child is sort of like a comment; a parenthetical statement or a footnote. You can continue the main thrust of an argument in sibling nodes -- just like writing an ordinary sequence of paragraphs in a word processor.

You can also organize things under headings. This is especially true if you wrote a sketchy outline first and then filled it in, or, if you have a lot of material in Workflowy and had to organize it. The "upper ontology" of my workflowy is mostly title-like,

single words or short noun phrases. As you get down in, bullets start to be sentences and paragraphs more often.

Obviously, all of this can be applied to Zettelkasten to some extent. The biggest difference is that “upper-level” cards are less likely to *just* be category titles; and, you can’t really organize things into nice categories after-the-fact because the addresses in Zettelkasten are fixed -- you can’t change them without breaking links. You can use redirect cards if you want to reorganize things, actually, but I haven’t done that very much in practice. Something which *has* worked for me to some extent is to reorganize things in the indexes. Once an index is too much of a big flat list, you can cluster entries into subjects. This new listing can be added as a child to the previous index, keeping the historical record; or, possibly, replace the old index outright. I discuss this more in the section on indexing.

Building Up Ideas over Long Time Periods

My idea books let me build up ideas over time to a greater extent than my peers who didn’t keep similar journals. However, because the linear format forces you to switch topics in a serial manner and “start over” when you want to resume a subject, you’re mostly restricted to what you can keep in your head. Your notebooks are a form of information storage, and you *can* go back and re-read things, but only if you remember the relevant item to go back and re-read.

Workflowy allowed me to build up ideas to a greater degree, incrementally adding thoughts until cascades of understanding changed my overall view.

Placing a New Idea

Because you’ve got all your ideas in one big outline, you can add in little ideas easily. Workflowy was easy enough to access via my smartphone (though they didn’t have a proper app at the time), so I could jot down an idea as I was walking to class, waiting for the bus, etc. I could easily navigate to the right location, at least, if I had organized the overall structure of the outline well. Writing one little idea would usually get more flowing, and I would add several points in the same location on the tree, or in nearby locations.

This idea of jotting down ideas while you’re out and about is very important. If you feel you don’t have enough ideas (be it for research, for writing fiction, for art -- whatever) my first question would be whether you have a good way to jot down little ideas as they occur to you.

The fact that you’re *forced* to *somewhat* fit all ideas into one big tree is also important. It makes you organize things in ways that are likely to be useful to you later.

Organizing Over Time

The second really nice thing workflowy did was allow me to go back and reorganize all the little ideas I had jotted down. When I sat down at a computer, I could take a look at my tree overall and see how well the categorization fit. This mostly took the form of

small improvements to the tree structure over time. Eventually, a cascade of small fixes turned into a major reorganization. At that point, I felt I had really learned something -- all the incremental progress built up into an overall shift in my understanding.

Again, this isn't really possible in paper-based Zettelkasten -- the address system is fixed. However, as I mentioned before, I've had some success doing this kind of reorganization within the indexes. It doesn't matter that the addresses of the cards are fixed if the way you actually *find* those addresses is mutable.

Limitations of Workflowy

Eventually, I noticed that I had a big pile of ideas which I hadn't really developed. I was jotting down ideas, sure. I was fitting them into an increasingly cohesive overall picture, sure. But I wasn't *doing anything* with them. I wasn't writing pages and pages of details and critique.

It was around this time that I realized I had gone more than three years without using a paper notebook very significantly. I started writing on paper again. I realized that there were all these habits of thinking which were tied to paper for me, and which I didn't really access if I didn't have a nice notebook and a nice pen -- the force of the long-practiced associations. It was like waking up intellectually after having gone to sleep for a long time. I started to *remember highschool*. It was a weird time. Anyway...

Dynalist

The next thing I tried was Dynalist.

The main advantage of Dynalist over Workflowy is that it takes a feature-rich rather than minimalistic approach. I like the clean aesthetics of Workflowy, but... eventually, there'll be some critical feature Workflowy just doesn't provide, and you'll want to make the jump to Dynalist. I use hardly any of the extra features of Dynalist, but the ones I *do* use, I *need*. For me, it's mostly the LaTeX support.

Another thing about Dynalist which felt very different for me was the file system. Workflowy forces you to keep everything in one big outline. Dynalist lets you create many outlines, which it treats as different files; and, you can organize them into folders (recursively). Technically, that's just another tree structure. In terms of UI, though, it made navigation much easier (because you can easily access a desired file through the file pane). Psychologically, it made me much more willing to start fresh outlines rather than add to one big one. This was both good and bad. It meant my ideas were less anchored in one big tree, but it eventually resulted in a big, disorganized pile of notes.

I did learn my lesson from Workflowy, though, and set things up in my Dynalist such that I actually developed ideas, rather than just collecting scraps forever.

Temporary Notes vs Organized Notes

I organized my Dynalist files as follows:

- A “log” file, in which I could write whatever I was thinking about. This was organized by date, although I would often go back and elaborate on things from previous dates.
- A “todo” file, where I put links to items inside “log” which I specifically wanted to go back and think more about. I would periodically sort the todo items to reflect my priorities. This gave me a list of important topics to draw from whenever I wasn’t sure what I wanted to think about.
- A bunch of other disorganized files.

This system wasn’t *great*, but it was a whole lot better at *actually developing ideas* than the way I kept things organized in Workflowy. I had realized that locking everything into a unified tree structure, while good for the purpose of slowly improving a large ontology which organized a lot of little thoughts, was keeping me from just writing whatever I was thinking about.

Dan Sheffler (whose essays I’ve already cited several times in this writeup) writes about realizing that his note-taking system was simultaneously trying to implement [two different goals](#): an organized long-term memory store, and “engagement notes” which are written to clarify thinking and have a more stream-of-consciousness style. My “log” file was essentially engagement notes, and my “todo” file was the long-term memory store.

For some people, I think an *essential part* of Zettelkasten is the distinction between temporary and permanent notes. Temporary notes are the disorganized stream-of-consciousness notes which Sheffler calls engagement notes. Temporary notes can also include all sorts of other things, such as todo lists which you make at the start of the day (and which only apply to that day), shopping lists, etc. Temporary notes can be kept in a linear format, like a notebook. Periodically, you review the temporary notes, putting the important things into Zettelkasten.

In *Taking Smart Notes*, Luhmann is described as transferring the important thoughts from the day into Zettel every evening. Sheffler, on the other hand, keeps a gap of at least 24 hours between taking down engagement notes and deciding what belongs in the long-term store. A gap of time allows the initial excitement over an idea to pass, so that only the things which still seem important the next day get into long-term notes. He also points out that this system enforces a small amount of spaced repetition, making it more likely that content is recalled later.

As for myself, I mostly write directly into my Zettelkasten, and I think it’s pretty great. However, I do find this to be difficult/impossible when taking quick notes during a conversation or a talk – when I try, then the resulting content in my Zettelkasten seems pretty useless (ie, I don’t come back to it and further develop those thoughts). So, I’ve started to carry a notebook again for those temporary notes.

I currently think of things like this:



Jots

These are the sort of small pointers to ideas which you can write down while walking, waiting for the bus, etc. The idea is stated very simply -- perhaps in a single word or a short phrase. A sentence at most. You might forget what it means after a week, especially if you don't record the context well. The first thing to realize about jots is to capture them at all, as already discussed. The second thing is to capture them in a place where you will be able to develop them later. I used to carry around a small pocket notebook for jots, after I stopped using Workflowy regularly. My plan was to review the jots whenever I filled a notebook, putting them in more long-term storage. This never happened: when I filled up a notebook, unpacking all the jots into something meaningful just seemed like too huge a task. It works better for me to jot things into permanent storage directly, as I did with Workflowy. I procrastinate too much on turning temporary notes into long term notes, and the temporary notes become meaningless.

Glosses

A gloss is a paragraph explaining the point of a jot. If a jot is the title of a Zettelkasten card, a gloss is the first paragraph (often written in a distinct color). This gives enough of an idea that the thought will not be lost if it is left for a few weeks (perhaps even years, depending). Writing a gloss is usually easy, and doing so is often enough to get the ideas flowing.

Development

This is the kind of writing I described in the 'notebooks' section. An idea is fleshed out. This kind of writing is often still comprehensible years later, although it isn't guaranteed to be.

Refinement

This is the kind of writing which is publishable. It nails the idea down. There's not really any end to this -- you can imagine expanding something from a blog post, to an academic paper, to a book, and further, with increasing levels of detail, gentle exposition, formal rigor -- but to a first approximation, anyway, you've eliminated all the contradictions, stated the motivating context accurately, etc.

I called the last item "refinement" rather than "communication" because, really, you can communicate your ideas at any of these stages. If someone shares a lot of context with you, they can understand your jots. That's really difficult, though. More likely, a research partner will understand your glosses. Development will be understandable to someone a little more distant, and so on.

At Long Last, Zettelkasten

I've been hammering home the idea of "linear" vs "nonlinear" formats as one of the big advantages of Zettelkasten. But workflowy and dynalist both allow nonlinear writing. Why should you be interested in Zettelkasten? Is it anything more than a way to implement workflowy-like writing for a paper format?

I've said that (at least for me) there's something extra-good about Zettelkasten which I don't really understand. But, there are a couple of important elements which make Zettelkasten more than just paper workflowy.

- **Hierarchy Plus Cross-Links:** A repeated theme across knowledge formats, including wikipedia and textbooks, is that you want both a hierarchical organization which makes it easy to get an overview and find things, and also a "cross-reference" type capability which allows related content to be linked -- creating a heterarchical web. I mentioned at the beginning that Zettelkasten forced me to create cross-links much more than I otherwise would, due to the use of small note-cards. Workflowy has "hierarchy" down, but it has somewhat poor "cross-link" capability. It has tags, but a tag system is not as powerful as hypertext. Because you *can* link to individual nodes, it's possible to use hypertext cross-links -- but the process is awkward, since you have to get the link to the node you want. Dynalist is significantly better in this respect -- it has an easy way to create a link to anything by searching for it (without leaving the spot you're at). But it lacks the wiki-style "magic link" capability, creating a new page when you make a link which has no target. Roam, however, provides this feature.
- **Atomicity:** The idea of creating pages organized around a single idea (again, an idea related to wikis). This is possible in Dynalist, but Zettelkasten practically forces it upon you, which for me was really good. Again, Roam manages to encourage this style.

Zettelkasten, Part 2: Further Advice

Card Layout

Don't stress about card formatting. You should write however feels natural to you. However, I thought you might like to see an example of what my cards tend to look like:



I'm left handed, so you may want to flip all of this around if you're right handed. I use the ring binder "backwards" from the intended configuration (the punched hole would usually be on the left, rather than the right). Also, I prefer portrait rather than landscape. Most people prefer to use 3x5 cards in landscape, I suppose.

Anyway, not every card will look exactly like the above. A card might just contain a bunch of free-writing, with no bulleted list. Or it might only contain a bulleted list, with no blurb at the beginning. Whatever works. I think my layout is close to Luhmann's and close to common advice -- but if you try to copy it religiously, you'll probably feel like Zettelkasten is awkward and restrictive.

The only absolutely necessary thing is the address. The address is the first thing you write on a new card. You don't ever want a card to go without an address. And it should be in a standard location, so that it is really easy to look through a bunch of cards for one with a specific address.

Don't feel bad if you start a card and leave it mostly blank forever. Maybe you thought you were going to elaborate an idea, so you made a new card, but it's got nothing but an address. That's ok. Maybe you will fill it later. Maybe you won't. Don't worry about it.

Mostly, a thought is continued through elaboration on bullet points. I might write something like "cont. 1.1a1a" at the bottom of the card if there's another card that's really a direct continuation, though. (Actually, I don't write "cont."; I just write the down arrow, which means the same thing.) If so, I'd write "see 1.1a1" in the upper left hand corner, to indicate that 1.1a1a probably doesn't make much sense on its own without consulting 1.1a1 -- moreso than usual for child cards. (Actually, I'd write another down arrow rather than "see", mirroring the down arrow on the previous card -- this indicates the direct-continuation relationship.)

In the illustration, I wrote links [in square brackets]. The truth is, I often put them in full rectangular boxes (to make them stand out more), although not always. Sometimes I put them in parentheses when I'm using them more as a noun, as in: "I think pizza (12a) might be relevant to pasta. [14x5b]" In that example, "(12a)" is the card for pizza. "[14x5b]" is a card continuing the whole thought "pizza might be relevant to pasta". So parentheses-vs-box is sort of like top-corner-vs-bottom, but for an individual line rather than a whole card.

Use of Color

The colors are true to my writing as well. For a long time, I wanted to try writing with multi-color click pens, because I knew some people found them very useful; but, I was unable to find any which satisfied my (exceptionally picky) taste. I don't generally go for ball-point pens; they aren't smooth enough. I prefer to write with felt-tip drawing pens or similar. I also prefer very fine tips (as a consequence of preferring my writing to be very small, as I mentioned previously) -- although I've also found that the appropriate line width varies with my mental state and with the subject matter. Fine lines are better for fine details, and for energetic mental states; broad lines are better for loose free-association and brainstorming, and for tired mental states.

In any case, a friend recommended the Hi-Tec C Coleto, a multi-color click pen which feels as smooth as felt-tip pens usually do (almost). You can buy whatever colors you

want, and they're available in a variety of line-widths, so you can customize it quite a bit.

At first I just used different colors haphazardly. I figured I would eventually settle on meanings for colors, if I just used whatever felt appropriate and experimented. Mostly, that meant that I switched colors to indicate a change of topic, or used a different color when I went back and annotated something (which really helps readability, by the way -- black writing with a bunch of black annotations scribbled next to it or between lines is hard to read, compared to purple writing with orange annotations, or whatever!). When I switched to Zettelkasten, though, I got more systematic with my use of color.

I roughly follow Lion Kimbro's advice about colors, from *How to Make a Complete Map of Every Thought you Think*:

Now lets talk about color.

Your pen has four colors: Red, Green, Blue, and Black

You will want to connect meaning with each color.

Here's my associations:

RED: Error, Warning, Correction

BLUE: Structure, Diagram, Picture, Links, Keys (in key-value pairs)

GREEN: Meta, Definition, Naming, Brief Annotation, Glyphs

BLACK: Main Content

I also use green to clarify sloppy writing later on. Blue is for Keys, Black is for values.

I hope that's self-explanatory.

If you make a correction, put it in red. Page numbers are blue. If you draw a diagram, make it blue. Main content in black.

Suppose you make a diagram: Start with a big blue box. Put the diagram in the box. (Or the other way around- make the diagram, than the box around it.) Put some highlighted content in black. Want to define a word? Use a green callout. Oops- there's a problem in the drawing- X it out in red, followed by the correction, in red.

Some times, I use black and blue to alternate emphasis. Black and blue are the easiest to see.

If I'm annotating some text in the future, and the text is black, I'll switch to using blue for content. Or vise versa.

Some annotations are red, if they are major corrections.

Always remember: Tolerate errors. If your black has run out, and you don't want to get up right away to fetch your backup pen, then just switch to blue. When the thoughts out, go get your backup pen.

The only big differences are that I use brown instead of black in my pen, I tend to use red for titles so that they stand out very clearly, and I use green for links rather than blue. (Honestly I think my use of green for links might be a mistake. You want links to stand out more.)

Index & Bibliography?

Bibliography

Taking Smart Notes describes two other kinds of cards: indexes, and bibliographical notes. I haven't made those work for me very effectively, however. Luhmann, the inventor of Zettelkasten, is described inventing Zettelkasten *as a way to organize notes originally made while reading*. I don't use it like that -- I mainly use it for organizing notes I make while thinking. So bibliography isn't of primary importance for me.

(Apparently Umberto Eco [similarly advises](#) keeping idea notes and reading notes on separate sets of index cards.)

Indexing

So I don't miss the bibliography cards. (Maybe I will eventually.) On the other hand, I definitely need some sort of index, but I'm not sure about the best way to keep it up to date. I only notice that I need it when I go looking for a particular card and it is difficult to find! When that happens, and I eventually find the card I wanted, I can jot down its address in an index. But, it would be nice to somehow avoid this. So, I've experimented with some ideas. Here are [someone else's thoughts on indexing](#) (for a digital zettelkasten).

Listing Assorted Cards

The first type of index which I tried lists "important" cards (cards which I refer to often). I just have one of these right now. The idea is that you write a card's name and address on this index if you find that you've had difficulty locating a card and wished it had been listed in your index. This sounds like it should be better than a simple list of the top-level numbered cards, since (as I mentioned earlier) cards like 11a often turn out to be more important than cards like 11. Unfortunately, I've found this not to be the case. The problem is that this kind of index is too hard to maintain. If I've just been struggling to find a card, my working memory is probably already over-taxed with all the stuff I wanted to do after finding that card. So I forget to add it to the index.

Topic Index

Sometimes it also makes sense to just make a new top-level card on which you list everything which has to do with a particular category. I have only done this once so far. It seems like this is the main mode of indexing which other people use? But I don't like the idea that well.

Listing Sibling Cards

When a card has enough children that they're difficult to keep track of, I add a "zero" card before all the other children, and this works as an index. So, for example, card 2a might have children 2a1, 2a2, 2a3, ... 2a15. That's a lot to keep track of. So I add 2a0, which gets an entry for 2a1-2a15, and any new cards added under 2a. It can also get an entry for particularly important descendants; maybe 2a3a1 is extra important and gets an entry.

For cards like 2, whose children are alphabetical, you can't really use "zero" to go before all the other children. I use " λ " as the "alphabetical zero" -- I sort it as if it comes before all the other letters in the alphabet. So, card "1 λ " lists 1a, 1b, etc.

The most important index is the index at 0, ei, the index of all top-level numbered cards. As I describe in the "card layout" section, a card already mostly lists its own children -- meaning that you don't need to add a new card to serve this purpose until things get unwieldy. However, top-level cards have no parents to keep track of them! So, you probably want an "absolute zero" card right away.

These "zero" cards also make it easier to keep track of whether a card with a particular address has been created yet. Every time you make a card, you add it to the appropriate zero card; so, you can see right away what the next address available is. This isn't the case otherwise, especially if your cards aren't currently sorted.

Kimbro's Mind Mapping

I've experimented with adapting Lion Kimbro's system from *How to Make a Complete Map of Every Thought You Think*. After all, a complete map of every thought you think sounds like the perfect index!

In my terminology, Lion Kimbro keeps only *jots* -- he was focusing on collecting and mapping, rather than developing, ideas. Jots were collected into topics and sub-topics. When an area accumulated enough jots, he would start a [mind map](#) for it. I won't go into all his specific mapping tips (although they're relevant), but basically, imagine putting the addresses of cards into clusters (on a new blank card) and then writing "anchor words" describing the clusters.

You built your tree in an initially "top-down" fashion, expanding trees by adding increasingly-nested cards. You're going to build the map "bottom-up": when a sub-tree you're interested in feels too large to quickly grasp, start a map. Let's say you're mapping card 8b4. You might already have an index of children at 8b4 λ ; if that's the case, you can start with that. Also look through all the descendants of 8b4 and pick out whichever seem most important. (If this is too hard, start by making maps for 8b4's *children*, and return to mapping 8b4 later.) Draw a new mind map, and place it at 8b4 λ a -- it is part of the index; you want to find it easily when looking at the index.

Now, the important thing is that *when you make a map for 8b*, you can take a look at the map for 8b4, as well as any maps possessed by other children of 8b. This means that you don't have to go through all of the descendants of 8b (which is good,

because there could be a lot). You just look at the maps, which already give you an overview. The map for 8b is going to take *the most important-seeming elements* from all of those sub-maps.

This allows important things to trickle up to the top. When you make a map at 0, you'll be getting all the most important stuff from deep sub-trees just by looking at the maps for each top-level numbered card.

The categories which emerge from mapping like this can be completely different from the concepts which initially seeded your top-level cards. You can make new top-level cards which correspond to these categories if you want. (I haven't done this.)

Now, when you're looking for something, you start at your top-level map. You look at the clusters and likely have some expectation about where it is (if the address isn't somewhere on your top-level map already). You follow the addresses to further maps, which give further clusters of addresses, until you land in a tree which is small enough to navigate without maps.

I've described all of this as if it's a one-time operation, but of course you keep adding to these maps, and re-draw updated maps when things don't fit well any more. If a map lives at 8b40a, then the updated maps can be 8b40b, 8b40c, and so on. You can keep the old maps around as a historical record of your shifting conceptual clusters.

Keeping Multiple Zettelkasten

A note system like Zettelkasten (or workflowy, dynalist, evernote, etc) is supposed to stick with you for years, growing with you and becoming a repository for your ideas. It's a big commitment.

It's difficult to optimize note-taking if you think of it that way, though. You can't experiment if you have to look before you leap. I would have never tried Zettelkasten if I thought I was committing to try it as my "next system" -- I didn't think it would work.

Similarly, I can't optimize my Zettelkasten very well with that attitude. A Zettelkasten is supposed to be one repository for everything -- you're not supposed to start a new one for a new project, for example. But, I have several Zettelkasten, to test out different formats: different sizes of card, different binders. It *is* still difficult to give alternatives a fair shake, because my two main Zettelkasten have built up momentum due to the content I keep in them.

I use a system of capital letters to cross reference between my Zettelkasten. For example, my main 3x5 Zettelkasten is "S" (for "small"). I have another Zettelkasten which is "M", and also an "L". When referencing card 1.1a within S, I just call it 1.1a. If I want to refer to it from a card in M, I call it S1.1a instead. And so on.

Apparently Luhmann [did something similar](#), starting a new Zettelkasten which occasionally referred to his first.

However, keeping multiple Zettelkasten for special topics is not necessarily a good idea. [Beware fixed categories](#). The danger is that categories limit what you write, or, become less appropriate over time. I've tried special-topic notebooks in the past, and

while it does sometimes work, I often end up conflicted about where to put something. (Granted, I have a similar conflict about where to put things in my several omni-topic Zettelkasten, but mostly the 3x5 system I've described here has won out -- for now.)

On the other hand, I suspect it's fine to create special topic zettelkasten for "very different" things. Creating a new zettelkasten because you're writing a new book is *probably* bad -- although it'll work fine for the goal of organizing material for writing books, it means your next book idea isn't coming from Zettelkasten. (Zettelkasten should contain/extend the thought process which generates book ideas in the first place, and it can't do that very well if you have to have a specific book idea in order to start a zettelkasten about it.) On the other hand, I suspect it is OK to keep a separate Zettelkasten for fictional creative writing. Factual ideas can spark ideas for fiction, but, the two are sufficiently different "modes" that it may make sense to keep them in physically separate collections.

The idea of using an extended address system to make references between multiple Zettelkasten can also be applied to address other things, outside of your Zettelkasten. For example, you might want come up with a way of adding addresses to your old notebooks so that you can refer to them easily. (For example, "notebook number: page number" could work.)

Should You Transfer Old Notes Into Zettelkasten?

Relatedly, since Zettelkasten ideally becomes a library of all the things you have been thinking about, it might be tempting to try and transfer everything from your existing notes into Zettelkasten.

(A lot of readers may not even be tempted to do this, given the amount of work it would take. Yet, those more serious about note systems might think this is a good idea -- or, might be too afraid to try Zettelkasten because they think they'd have to do this.)

I think transferring older stuff into Zettelkasten can be useful, but, trying to make it happen right away as one big project is most likely not worth it.

- It's true that part of the usefulness of Zettelkasten is the interconnected web of ideas which builds up over time, and the "high-surface-area" format which makes it easy to branch off any part. However, not all the payoff is long-term: it should also be useful *in the moment*. You're not *only* writing notes because they may help you develop ideas in the future; the act of writing the notes should be helping you develop ideas now.
- You should probably only spend time putting ideas into Zettelkasten if you're excited about further developing those ideas right now. You should not just be copying over ideas into Zettelkasten. You should be improving ideas, thinking about where to place them in your address hierarchy, interlinking them with other ideas in your Zettelkasten via address links, and taking notes on any new ideas sparked by this process. Trying to put all your old notes into Zettelkasten at once will likely make you feel hurried and unwilling to develop things further

as you go. This will result in a pile of mediocre notes which will ultimately be less useful.

- I mentioned the breadth-first vs depth-first distinction earlier. Putting all of your old notes into Zettelkasten is an extremely breadth-first strategy, which likely doesn't give you enough time to go deep into further developing any one idea.

What about the dream of having all your notes in one beautiful format? Well, it is true that old notes in different formats may be harder to find, since you have to remember what format the note you want was written in, or check all your old note systems to find the note you want. I think it just isn't worth the cost to fix this problem, though, especially since you should probably try many different systems to find a good one that works for you, and you can't very well port all your notes to each new system.

Zettelkasten should be *an overall improvement compared to a normal notebook* -- if it isn't, you have no business using it. Adding a huge up-front cost of transferring notes undermines that. Just pick Zettelkasten up when you want to use it to develop ideas further.

Depth-first vs Breadth-first

Speaking of depth-first vs breadth-first, how should you balance those two modes?

Luckily, this problem has some relevant computer science theory behind it. I tend to think of it in terms of iterative-deepening A* heuristic search (IDA*).

The basic idea is this: the advantage of depth-first search is that you can minimize memory cost by only maintaining the information related to the path you are currently trying. However, depth-first search can easily get stuck down a fruitless path, while breadth-first search has better guarantees. IDA* balances the two approaches by going depth-first, but giving up when you get too deep, backing up, and trying a new path. (The A* aspect is that you define "too deep" in a way which also depends on how promising a path seems, based on an optimistic assessment.) This way, you simulate a breadth-first search by a series of depth-first sprints. This lets you focus your attention on a small set of ideas at one time.

Once you've explored all the paths to a certain level, your tolerance defining "too deep" increases, and you start again. You can think of this as becoming increasingly willing to spend a lot of time going down difficult technical paths as you confirm that easier options don't exist.

Of course, this isn't a perfect model of what you should do. But, it seems to me that a note-taking system should aspire to support and encourage something resembling this. More generally, I want to get across the idea of thinking of your existing research methodology as an algorithm (possibly a bad one), and trying to think about how it could be improved. Don't try to force yourself to use any particular algorithm just because you think you should; but, if you can find ways to nudge yourself toward more effective algorithms, that's probably a good idea.

Inventing Shorthand/Symbology

I don't think writing speed is a big bottleneck to thinking speed. Even though I "think by writing", a lot of my time is spent... well... thinking. However, once I know what I want to write, writing *does* take time. When inspiration really strikes, I might know more or less what I want to say several paragraphs ahead of where I've actually written to. At times like that, it seems like every second counts -- the faster I write, the more ideas I get down, the less I forget before I get to it.

So, it seems worth putting *some* effort into writing faster. (Computer typing is obviously a thing to consider here, too.) Shorthand, and special symbols, are something to try.

There's also the issue of space. I know I advocate for small cards, which are intentionally limiting space. But you don't want to waste space if you don't have to. The point is to comprehend as much as possible as easily as possible. Writing bullet points and using indentation to make outlines is an improvement over traditional paragraphs because it lets you see more at a glance. Similarly, using abbreviations and special symbols will improve this.

I've tried several times to learn "proper" shorthand. Maybe I just haven't tried hard enough, but it seems like basically all shorthand systems work by leaving out information. Once you're used to them, they're easy enough to read shortly after you've written them -- when you still remember more or less what they said. However, they don't actually convey enough information to fully recover what was written if you don't have such a guess. Basically, they don't improve readability. They compress things down to the point where they're hard to decipher, for the sake of getting as much speed as possible.

On the other hand, I've spent time experimenting with changes to my own handwriting which improve speed without compromising readability. Pay attention to what takes you the most time to write, and think about ways to streamline that.

Lion Kimbro emphasizes that you come up with ways to abbreviate things you commonly repeat. He describes using the Japanese symbols for days of the week and other common things in his system. The Bullet Journaling community has created its own symbology. Personally, I've experimented with a variety of different reference symbols which mean different sorts of things (beyond the () vs [] distinction I've mentioned).

The Bullet Journaling community has thought a lot about short-and-fast writing for the purpose of getting things out quickly and leaving more space on the page. They also have their own symbology which may be worth taking a look at. (I don't yet use it, but I may switch to it or something similar eventually.)

Well, that's all I want to say for now. I may add to this document in the future. For now, best of luck developing ideas!

Follow-Up Reports

June 12, 2020. My index-card Zettelkasten has not continued to grow at the rate it did initially. However, I still find the address system of Zettelkasten almost

indispensable for note-taking. (I would say "I can't imagine what I did before" except that I know exactly what I did before; it was just way worse.) I have several different Zettelkasten in different formats, which I switch between.

One of the main ways I now use Zettelkasten is in notebooks with fixed-order pages. This means I **can never sort the cards**; pages remain in the order which they're originally written in. However, maintaining the address system is still *extremely* useful, and allows me to have thoughts which I would not otherwise be capable of having. As I mentioned, I don't sort my index cards that often anyway, and I feel like I lose important recency information when I do so. So losing the ability to rearrange pages was not so bad! Although, I still miss it sometimes and do still use other formats which allow me to rearrange pages appropriately.

I still mainly find Zettelkasten useful on the order of minutes, weeks, and months. After that long, my notes start to become "stale" and I'm better off starting fresh rather than continuing to build on my old note structures about a particular idea. (Old notes are still useful for reference purposes, but I very rarely add to them to further develop the ideas therein, preferring to start fresh with a new tree of ideas.)

This is partly because the Zettelkasten note structure is not very rearrangeable, and so I can't "fix" old structures to reflect new thinking. As I've discussed, I've had better luck with digital note structures continuing to have real use over long time periods. But it's not just that -- I've still *mainly* experienced note structures "going stale" in digital formats, as well.

There are two possible responses to this that I see. (1) Try to develop better techniques to make long-term useful structures. (2) Embrace the tendency, and optimize for the more immediate usefulness. I've tended toward (2). I worry that optimizing for (1) too much would create only hypothetical value at the cost of real value -- like when you delay gratification indefinitely.

Relatedly, I still haven't made much use of indexing.

The unexpected difficulty of comparing AlphaStar to humans

This is crossposted from the [AI Impacts blog](#).

Artificial intelligence defeated a pair of professional Starcraft II players for the first time in December 2018. Although this was generally regarded as an impressive achievement, it quickly became clear that not everybody was satisfied with how the AI agent, called AlphaStar, interacted with the game, or how its creator, DeepMind, presented it. Many observers complained that, in spite of DeepMind's claims that it performed at similar speeds to humans, AlphaStar was able to control the game with greater speed and accuracy than any human, and that this was the reason why it prevailed.

Although I think this story is mostly correct, I think it is harder than it looks to compare AlphaStar's interaction with the game to that of humans, and to determine to what extent this mattered for the outcome of the matches. Merely comparing raw numbers for actions taken per minute (the usual metric for a player's speed) does not tell the whole story, and appropriately taking into account mouse accuracy, the differences between combat actions and non-combat actions, and the control of the game's "camera" turns out to be quite difficult.

Here, I begin with an overview of Starcraft II as a platform for AI research, a timeline of events leading up to AlphaStar's success, and a brief description of how AlphaStar works. Next, I explain why measuring performance in Starcraft II is hard, show some analysis on the speed of both human and AI players, and offer some preliminary conclusions on how AlphaStar's speed compares to humans. After this, I discuss the differences in how humans and AlphaStar "see" the game and the impact this has on performance. Finally, I give an update on DeepMind's current experiments with Starcraft II and explain why I expect we will encounter similar difficulties when comparing human and AI performance in the future.

Why Starcraft is a Target for AI Research

Starcraft II has been a target for AI for several years, and some readers will recall that Starcraft II appeared on our [2016 expert survey](#). But there are many games and many AIs that play them, so it may not be obvious why Starcraft II is a target for research or why it is of interest to those of us that are trying to understand what is happening with AI.

For the most part, Starcraft II was chosen because it is popular, and it is difficult for AI. Starcraft II is a real time strategy game, and like similar games, it requires a variety of tasks: harvesting resources, constructing bases, researching technology, building armies, and attempting to destroy their opponent's base are all part of the game. Playing it well requires balancing attention between many things at once: planning ahead, ensuring that one's units¹ are good counters for the enemy's units, predicting opponents' moves, and changing plans in response to new information. There are other aspects that make it difficult for AI in particular: it has imperfect information², an extremely large action space, and takes place in real time. When humans play, they engage in long term planning, making the best use of their limited capacity for attention, and crafting ploys to deceive the other players.

The game's popularity is important because it makes it a good source of extremely high human talent and increases the number of people that will intuitively understand how difficult the task is for a computer. Additionally, as a game that is designed to be suitable for high-level competition, the game is carefully balanced so that competition is fair, does not favor just one strategy³, and does not rely too heavily on luck.

Timeline of Events

To put AlphaStar's performance in context, it helps to understand the timeline of events over the past few years:

November 2016: Blizzard and DeepMind [announce](#) they are launching a new project in Starcraft II AI

August 2017: DeepMind [releases](#) the Starcraft II API, a set of tools for interfacing AI with the game

March 2018: Oriol Vinyals gives an [update](#), saying they're making progress, but he doesn't know if their agent will be able to beat the best human players

November 3, 2018: Oriol Vinyals gives another update at a Blizzcon panel, and shares a sequence of videos demonstrating AlphaStar's progress in learning the game, including leaning to win against the hardest built-in AI. When asked if they could play against it that day, he says "For us, it's still a bit early in the research."

December 12, 2018: AlphaStar wins five straight matches against TLO, a professional Starcraft II player, who was playing as Protoss⁴, which is off-race for him. DeepMind keeps the matches secret.

December 19, 2018: AlphaStar, given an additional week of training time⁵, wins five consecutive Protoss vs Protoss matches vs MaNa, a pro Starcraft II player who is higher ranked than TLO and specializes in Protoss. DeepMind continues to keep the victories a secret.

January 24, 2019: DeepMind [announces](#) the successful test matches vs TLO and MaNa in a live video feed. MaNa plays a live match against a version of AlphaStar which had more constraints on how it "saw" the map, forcing it to interact with the game in a way more similar to humans⁶. AlphaStar loses when MaNa finds a way to exploit a blatant failure of the AI to manage its units sensibly. The replays of all the matches are released, and people start arguing⁷ about how (un)fair the matches were, whether AlphaStar is any good at making decisions, and how honest DeepMind was in presenting the results of the matches.

July 10, 2019: DeepMind and Blizzard announce that they will allow an experimental version of AlphaStar to play on the European ladder⁸, for players who opt in. The agent will play anonymously, so that most players will not know that they are playing against a computer. Over the following weeks, players attempt to discern whether they played against the agent, and some post replays of matches in which they believe they were matched with the agent.

How AlphaStar works

The best place to learn about AlphaStar is from [DeepMind's page](#) about it. There are a few particular aspects of the AI that are worth keeping in mind:

It does not interact with the game like a human does: Humans interact with the game by looking at a screen, listening through headphones or speakers, and giving commands through a mouse and keyboard. AlphaStar is given a list of units or buildings and their attributes, which includes things like their location, how much damage they've taken, and which actions they're able to take, and gives commands directly, using coordinates and unit identifiers. For most of the matches, it had access to information about anything that wouldn't normally be hidden from a human player, without needing to control a "camera" that focuses on only one part of the map at a time. For the final match, it had a camera restriction similar to humans, though it still was not given screen pixels as input. Because it gives commands directly through the game, it does not need to use a mouse accurately or worry about tapping the wrong key by accident.

It is trained first by watching human matches, and then through self-play: The neural network is trained first on a large database of matches between humans, and then by playing against versions of itself.

It is a set of agents selected from a tournament: Hundreds of versions of the AI play against each other, and the ones that perform best are selected to play against human players. Each one has its own set of units that it is incentivized to use via reinforcement learning, so that they each play with different strategies. TLO and MaNa played against a total of 11 agents, all of which were selected from the same tournament, except the last one, which had been substantially modified. The agents that defeated MaNa had each played for hundreds of years in the virtual tournament⁹.

January/February Impressions Survey

Before deciding to focus my investigation on a comparison between human and AI performance in Starcraft II, I conducted an informal survey with my Facebook friends, my colleagues at AI Impacts, and a few people from an effective altruism Facebook group. I wanted to know what they were thinking about the matches in general, with an emphasis on which factors most contributed to the outcome of the matches. I've put details about my analysis and the full results of the survey in the appendix at the end of this article, but I'll summarize a few major results here.

Forecasts

The timing and nature of AlphaStar's success seems to have been mostly in line with people's expectations, at least at the time of the announcement. Some respondents did not expect to see it for a year or two, but on average, AlphaStar was less than a year earlier than expected. It is probable that some respondents had been expecting it to take longer, but updated their predictions in 2016 after finding out that DeepMind was working on it. For future expectations, a majority of respondents expect to see an agent (not necessarily AlphaStar) that can beat the best humans without any of the current caveats within two years. In general, I do not think that I worded the forecasting questions carefully enough to infer very much from the answers given by survey respondents.

Some readers may be wondering how these survey results compare to those of our more careful 2016 survey, or how we should view the earlier survey results in light of

MaNa and TLOs defeat at the hands of AlphaStar. The 2016 survey specified an agent that only receives a video of the screen, so that prediction has not yet resolved. But the median respondent assigned 50% probability of seeing such an agent that can defeat the top human players at least 50% of the time by 2021[10](#). I don't personally know how hard it is to add in that capability, but my impression from speaking to people with greater machine learning expertise than mine is that this is not out of reach, so these predictions still seem reasonable, and are not generally in disagreement with the results from my informal survey.

Speed

Nearly everyone thought that AlphaStar was able to give commands faster and more accurately than humans, and that this advantage was an important factor in the outcome of the matches. I looked into this in more detail, and wrote about it in the next section.

Camera

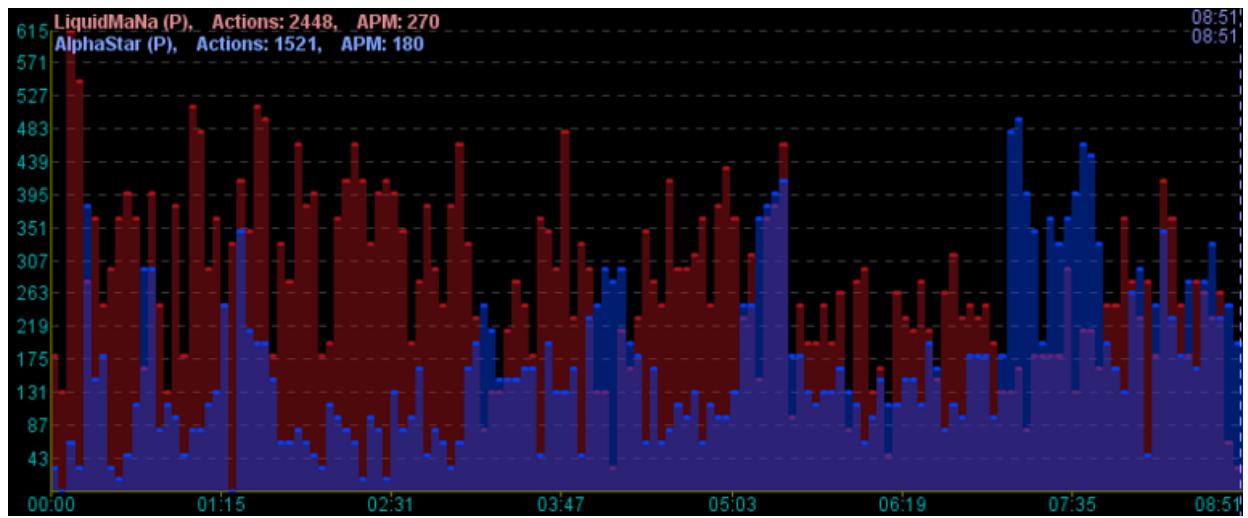
As I mentioned in the description of AlphaStar, it does not see the game the same way that humans do. Its visual field covered the entire map, though its vision was still affected by the usual fog of war[11](#). Survey respondents ranked this as an important factor in the outcome of the matches.

Given these results, I decided to look into the speed and camera issues in more detail.

The Speed Controversy

Starcraft is a game that rewards the ability to micromanage many things at once and give many commands in a short period of time. Players must simultaneously build their bases, manage resource collection, scout the map, research better technology, build individual units to create an army, and fight battles against other players. The combat is sufficiently fine grained that a player who is outnumbered or outgunned can often come out ahead by exerting better control over the units that make up their military forces, both on a group level and an individual level. For years, there have been simple Starcraft II bots that, although they cannot win a match against a highly-skilled human player, can do [amazing things](#) that humans can't do, by controlling dozens of units individually during combat. In practice, human players are limited by how many actions they can take in a given amount of time, usually measured in actions per minute (APM). Although DeepMind imposed restrictions on how quickly AlphaStar could react to the game and how many actions it could take in a given amount of time, many people believe that the agent was sometimes able to act with superhuman speed and precision.

Here is a graph[12](#) of the APM for MaNa (red) and AlphaStar (blue), through the second match, with five-second bins:



Actions per minute for MaNa (red) and AlphaStar (blue) in their second game. The horizontal axis is time, and the vertical axis is 5 second average APM.

At first glance, this looks reasonably even. AlphaStar has both a lower average APM (180 vs MaNa's 270) for the whole match, and a lower peak 5 second APM (495 vs Mana's 615). This seems consistent with DeepMind's claim that AlphaStar was restricted to human-level speed. But a more detailed look at which actions are actually taken during these peaks reveals some crucial differences. Here's a sample of actions taken by each player during their peaks:

| MaNa | | | | | AlphaStar | | | | |
|------|-------|------------|--------------------|------------------------|-----------|-------|-----------|----------------|--|
| I | Time | User | Event | Parameters | I | Time | User | Event | Parameters |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:43 | AlphaStar | SelectionDelta | Add Phoenix x10 (2b6,2sa,31i,3ch,5pp) |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:43 | AlphaStar | Cmd | [Right click] ; target unit: Zealot (tag=3r5 |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:43 | AlphaStar | SelectionDelta | Add Phoenix x10 (2b6,2sa,31i,3ch,5pp) |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:43 | AlphaStar | Cmd | Graviton Beam; target unit: Stalker (tag= |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:43 | AlphaStar | SelectionDelta | Add Stalker x4 (37i,382,3dd,3eq); Rem |
| 1 | 00:10 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:43 | AlphaStar | Cmd | Attack; target point: x=65.227, y=15.273 |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:43 | AlphaStar | SelectionDelta | Add Stalker x4 (37i,382,3dd,3eq); Rem |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:43 | AlphaStar | Cmd | [Right click] ; target point: x=62.711, y=1 |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:44 | AlphaStar | SelectionDelta | Add Phoenix x7 (2sa,31i,3ch,5pph,3pt, |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:44 | AlphaStar | SelectionDelta | Add Phoenix x7 (2sa,31i,3ch,5pph,3pt, |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:44 | AlphaStar | Cmd | Graviton Beam; target unit: Stalker (tag= |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 3 | 1 | 07:44 | AlphaStar | SelectionDelta | Add Phoenix x6 (3ch,5pph,3gu,3pt,3v5, |
| 1 | 00:11 | LiquidMaNa | ControlGroupUpdate | Select control group 1 | 1 | 07:44 | AlphaStar | Cmd | [Right click] ; target point: x=62.352, y=1 |
| | | | | | 1 | 07:44 | AlphaStar | SelectionDelta | Add Phoenix x6 (3ch,5pph,3gu,3pt,3v5, |

Lists of commands for MaNa and AlphaStar during each player's peak APM for game 2

MaNa hit his APM peaks early in the game by using hot keys to twitchily switch back and forth between control groups¹³ for his workers and the main building in his base. I don't know why he's doing this: maybe to warm up his fingers (which apparently is a thing), as a way to watch two things at once, to keep himself occupied during the slow parts of the early game, or some other reason understood only by the kinds of people that can produce Starcraft commands faster than I can type. But it drives up his peak APM, and probably is not very important to how the game unfolds¹⁴. Here's what MaNa's peak APM looked like at the beginning of Game 2 (if you look at the bottom of

the screen, you can see that the units he has selected switches back-and-forth between his workers and the building that he uses to make more workers):



MaNa's play during his peak APM for match 2. Most of his actions consist of switching between control groups without giving new commands to any units or buildings

AlphaStar hit peak APM in combat. The agent seems to reserve a substantial portion of its limited actions budget until the critical moment when it can cash them in to eliminate enemy forces and gain an advantage. Here's what that looked like near the end of game 2, when it won the engagement that probably won it the match (while still taking a few actions back at its base to keep its production going):



AlphaStar's play during its peak APM in match 2. Most of its actions are related to combat, and require precise timing.

It may be hard to see what exactly is happening here for people who have not played the game. AlphaStar (blue) is using extremely fine-grained control of its units to defeat MaNa's army (red) in an efficient way. This involves several different actions: Commanding units to move to different locations so they can make their way into his base while keeping them bunched up and avoiding spots that make them vulnerable, focusing fire on MaNa's units to eliminate the most vulnerable ones first, using special abilities to lift MaNa's units off the ground and disable them, and redirecting units to attack MaNa's workers once a majority of MaNa's military units are taken care of.

Given these differences between how MaNa and AlphaStar play, it seems clear that we can't just use raw match-wide APM to compare the two, which most people paying attention seem to have noticed fairly quickly after the matches. The more difficult question is whether AlphaStar won primarily by playing with a level of speed and accuracy that humans are incapable of, or by playing better in other ways. Though based on the analysis that I am about to present I think the answer is probably that AlphaStar won through speed, I also think the question is harder to answer definitively than many critics of DeepMind are making it out to be.

A [very fast human](#) can average well over 300 APM for several minutes, with 5 second bursts at over 600 APM. Although these bursts are not always throwaway commands like those from the MaNa vs AlphaStar matches, they tend not to be commands that require highly accurate clicking, or rapid movement across the map. Take, for example, this 10 second, 600 APM peak from current top player Serral:



Serral's play during a 10 second, 600 APM peak

Here, Serral has just finished focusing on a pair of battles with the other player, and is taking care of business in his base, while still picking up some pieces on the battlefield. It might not be obvious why he is issuing so many commands during this time, so let's look at the list of commands:

| I | Time | User | Event | Parameters |
|---|-------|--------|--------------------|---|
| | 12:23 | Serral | ControlGroupUpdate | Select control group 3 |
| | 12:24 | Serral | SelectionDelta | Add Larva x20 (5qv4,4ns,4wq,5rc1,5rci,53t,56x,5dj,5o2,5 |
| | 12:24 | Serral | Cmd | Morph to Hydralisk |
| | 12:24 | Serral | SelectionDelta | Add Zerg Cocoon (5qv4); Remove 1 unit |
| | 12:24 | Serral | SelectionDelta | Add Zerg Cocoon (4ns); Remove 1 unit |
| | 12:24 | Serral | Cmd | Morph to Roach |
| | 12:24 | Serral | SelectionDelta | Add Zerg Cocoon (4wq); Remove 1 unit |
| | 12:24 | Serral | SelectionDelta | Remove 17 units |
| | 12:24 | Serral | ControlGroupUpdate | Assign to control group 1 (add selection) |
| | 12:25 | Serral | ControlGroupUpdate | Select control group 1 |
| | 12:25 | Serral | Cmd | [Right click] ; target point: x=51.255, y=52.663, z=3.996 |
| | 12:25 | Serral | ControlGroupUpdate | Select control group 2 |
| | 12:25 | Serral | CameraUpdate | x=72.277; y=119.434 |
| | 12:25 | Serral | SelectionDelta | Remove 1 unit |

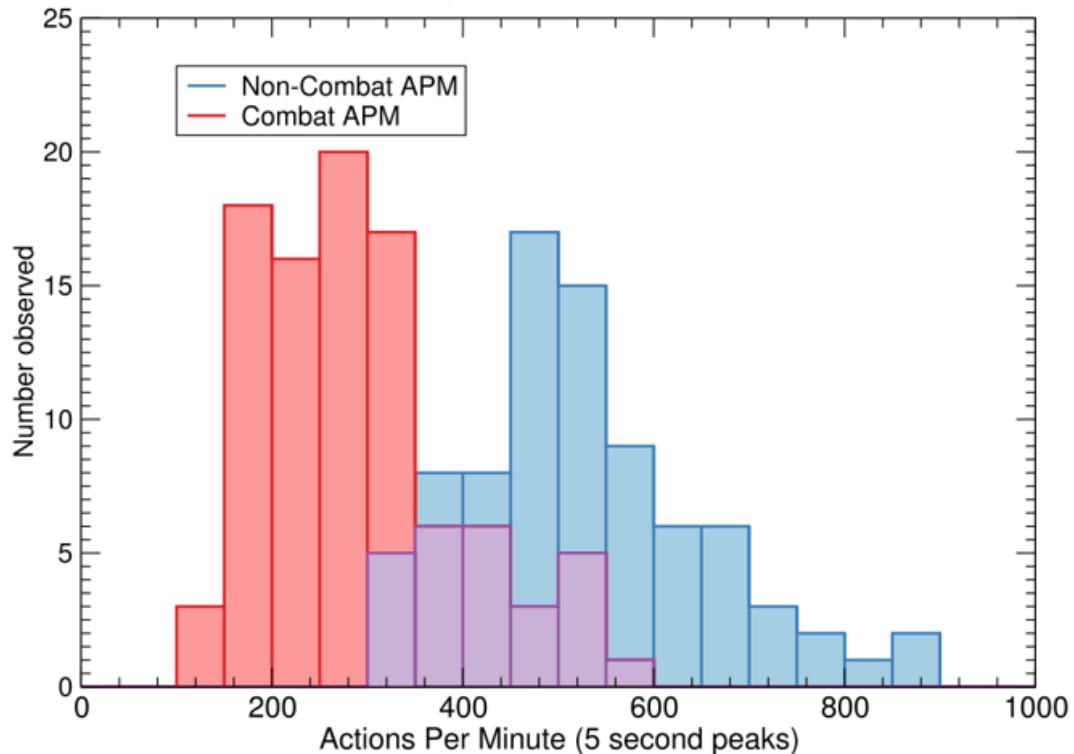
The lines that say “Morph to Hydralisk” and “Morph to Roach” represent a series of repeats of that command. For a human player, this is a matter of pressing the same hotkey many times, or even just holding down the key to give the command very rapidly¹⁵. You can see this in the gif by looking at the bottom center of the screen where he selects a bunch of worm-looking things and turns them all into a bunch of egg-looking things (it happens very quickly, so it can be easy to miss).

What Serral is doing here is difficult, and the ability to do it only comes with years of practice. But the raw numbers don’t tell the whole story. Taking 100 actions in 10 seconds is much easier when a third of those actions come from holding down a key for a few hundred milliseconds than when they each require a press of a different key or a precise mouse click. And this is without all the extraneous actions that humans often take (as we saw with MaNa).

Because it seems to be the case that peak human APM happens outside of combat, while AlphaStar’s wins happened during combat APM peaks, we need to do a more detailed analysis to determine the highest APM a human player can achieve during combat. To try to answer this question, I looked at approximately ten APM for each of the 5 games between AlphaStar and MaNa, as well as each of another 15 replays between professional Starcraft II players. The peaks were chosen so that roughly half were the largest peak at any time during the match and the rest were strictly during combat. My methodology for this is given in the appendix. Here are the results for just the human vs human matches:

APM for Professional Starcraft II Matches

177 peaks from 15 matches

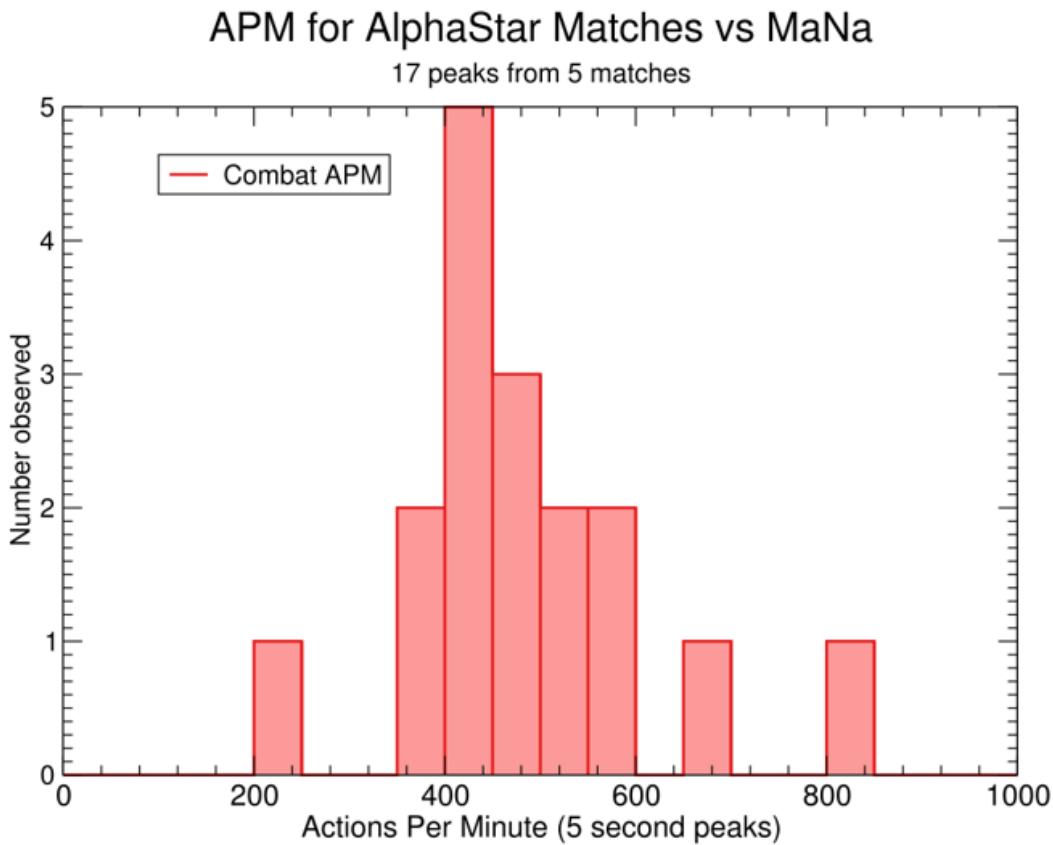


Histogram of 5-second APM peaks from analyzed matches between human professional players in a tournament setting. The blue bars are peaks achieved outside of combat, while the red bars are those achieved during combat.

Provisionally, it looks like pro players frequently hit approximately 550 to 600 APM outside of combat before the distribution starts to fall off, and they peak at around 200-350 during combat, with a long right tail. As I was doing this, however, I found that all of the highest APM peaks had one thing in common with each other that they did not have in common with all of the lower APM peaks, which is that it was difficult to tell when a player's actions are primarily combat-oriented commands, and when they are mixed in with bursts of commands for things like training units. In particular, I found that the combat situations with high APM tended to be similar to the Serral gif above, in that they involve spam clicking and actions related to the player's economy and production, which was probably driving up the numbers. I give more details in the appendix, but I don't think I can say with confidence that any players were achieving greater than 400-450 APM in combat, in the absence of spurious actions or macromanagement commands.

The more pertinent question might be what the lowest APM is that a player can have while still succeeding at the highest level. Since we know that humans can succeed without exceeding this APM, it is not an unreasonable limitation to put on AlphaStar. The lowest peak APM in combat I saw for a winning player in my analysis was 215, though it could be that I missed a higher peak during combat in that same match.

Here is a histogram of AlphaStar's combat APM:



The smallest 5-second APM that AlphaStar needed to win a match against MaNa was just shy of 500. I found 14 cases in which the agent was able to average over 400 APM for 5 seconds in combat, and six times when the agent averaged over 500 APM for more than 5 seconds. This was done with perfect accuracy and no spam clicking or control group switching, so I think we can safely say that its play was faster than is required for a human to win a match in a professional tournament. Given that I found no cases where a human was clearly achieving this speed in combat, I think I can comfortably say that AlphaStar had a large enough speed advantage over MaNa to have substantially influenced the match.

It's easy to get lost in numbers, so it's good to take a step back and remind ourselves of the insane level of skill required to play Starcraft II professionally. The top professional players already play with what looks to me like superhuman speed, precision, and multitasking, so it is not surprising that the agent that can beat them is so fast. Some observers, especially those in the Starcraft community, have indicated that they will not be impressed until AI can beat humans at Starcraft II at sub-human APM. There is some extent to which speed can make up for poor strategy and good strategy can make up for a lack of speed, but it is not clear what the limits are on this trade-off. It may be very difficult to make an agent that can beat professional Starcraft II players while restricting its speed to an undisputedly human or sub-human level, or it may simply be a matter of a couple more weeks of training time.

The Camera

As I explained earlier, the agent interacts with the game differently than humans. As with other games, humans look at a screen to know what's happening, use a mouse and keyboard to give commands, and need to move the game's 'camera' to see different parts of the play area. With the exception of the final exhibition match against MaNa, AlphaStar was able to see the entire map at once (though much of it is concealed by the fog of war most of the time), and had no need to select units to get information about them. It's unclear just how much of an advantage this was for the agent, but it seems likely that it was significant, if nothing else because it did not suffer from the APM overhead just to look around and get information from the game. Furthermore, seeing the entire map makes it easier to simultaneously control units across the map, which AlphaStar used to great effect in the first five matches against MaNa.

For the exhibition match in January, DeepMind trained a version of AlphaStar that had similar camera control to human players. Although the agent still saw the game in a way that was abstracted from the screen pixels that humans see, it only had access to about one screen's worth of information at a time, and it needed to spend actions to look at different parts of the map. A further disadvantage was that this version of the agent only had half as much training time as the agents that beat MaNa.

Here are three factors that may have contributed to AlphaStar's loss:

1. The agent was unable to deal effectively with the added complication of controlling the camera
2. The agent had insufficient training time
3. The agent had easily exploitable flaws the whole time, and MaNa figured out how to use them in match 6

For the third factor, I mean that the agent had sufficiently many exploitable flaws that were obvious enough to human players that any skilled human player could find at least one during a small number of games. The best humans do not have a sufficient number of such flaws to influence the game with any regularity. Matches in professional tournaments are not won by causing the other player to make the same obvious-to-humans mistake over and over again.

I suspect that AlphaStar's loss in January is mainly due to the first two factors. In support of 1, AlphaStar seemed less able to simultaneously deal with things happening on opposite sides of the map, and less willing to split its forces, which could plausibly be related to an inability to simultaneously look at distant parts of the map. It's not just that the agent had to move the camera to give commands on other parts of the map. The agent had to remember what was going on globally, rather than being able to see it all the time. In support of 2, the agent that MaNa defeated had only as much training time as the agents that went up against TLO, and those agents lost to the agents that defeated MaNa 94% of the time during training [16](#).

Still, it is hard to dismiss the third factor. One way in which an agent can improve through training is to encounter tactics that it has not seen before, so that it can react well if it sees it in the future. But the tactics that it encounters are only those that another agent employed, and without seeing the agents during training, it is hard to know if any of them learned the harassment tactics that MaNa used in game 6, so it is hard to know if the agents that defeated MaNa were susceptible to the exploit that he used to defeat the last agent. So far, the evidence from DeepMind's more recent experiment pitting AlphaStar against the broader Starcraft community (which I will go

into in the next section) suggests that the agents do not tend to learn defenses to these types of exploits, though it is hard to say if this is a general problem or just one associated with low training time or particular kinds of training data.

AlphaStar on the Ladder

For the past couple months, as of this writing, skilled European players have had the opportunity to play against AlphaStar as part of the usual system for matching players with those of similar skill. For the version of AlphaStar that plays on the European ladder, DeepMind claims to have made changes that address the camera and action speed complaints from the January matches. The agent needs to control the camera, and [they say](#) they have placed restrictions on AlphaStar's performance in consultation with pro players, particularly the maximum actions per minute and per second that the agent can make. I will be curious to see what numbers they arrive at for this. If this was done in an iterative way, such that pro players were allowed to see the agent play or to play against it, I expect they were able to arrive at a good constraint. Given the difficulty that I had with arriving at a good value for a combat APM restriction, I'm less confident that they would get a good value just by thinking about it, though if they were sufficiently conservative, they probably did alright.

Another reason to expect a realistic APM constraint is that DeepMind wanted to run the European ladder matches as a blind study, in which the human players did not know they were playing against an AI. If the agent were to play with the superhuman speed and accuracy that AlphaStar did in January, it would likely give it away and spoil the experiment.

Although it is unclear that any players were able to tell they were playing against an AI during their match, it does seem that some were able to figure it out after the fact. One example comes from Lowko, who is a Dutch player who streams and does commentary for games. During a stream of a ladder match in Starcraft II, he noticed the player was doing some strange things near the end of the match, like lifting their buildings[17](#) when the match had clearly been lost, and air-dropping workers into Lowko's base to kill units. Lowko did eventually win the match. Afterward, he was able to view the replay from the match and see that the player he had defeated did some very strange things throughout the entire match, the most notable of which was how the player controlled their units. The player used no control groups at all, which is, as far as I know, not something anybody does at high-level play[18](#). There were many other quirks, which he describes [in his entertaining video](#), which I highly recommend to anyone who is interested.

Other players have released replay files from matches against players they believed were AlphaStar, and they show the same lack of control groups. This is great, because it means we can get a sense of what the new APM restriction is on AlphaStar. There are now dozens of replay files from players who claim to have played against the AI. Although I have not done the level of analysis that I did with the matches in the APM section, it seems clear that they have drastically lowered the APM cap, with the matches I have looked at topping out at 380 APM peaks, which did not even occur in combat.

It seems to be the case that DeepMind has brought the agent's interaction with the game more in line with human capability, but we will probably need to wait until they release the details of the experiment before we can say for sure.

Another notable aspect of the matches that people are sharing is that their opponent will do strange things that human players, especially skilled human players almost never do, most of which are detrimental to their success. For example, they will construct buildings that block them into their own base, crowd their units into a dangerous bottleneck to get to a cleverly-placed enemy unit, and fail to change tactics when their current strategy is not working. These are all the types of flaws that are well-known to exist in game-playing AI going back to much older games, including the original Starcraft, and they are similar to the flaw that MaNa exploited to defeat AlphaStar in game 6.

All in all, the agents that humans are uncovering seem to be capable, but not superhuman. Early on, the accounts that were identified as likely candidates for being AlphaStar were winning about 90-95% of their matches on the ladder, achieving Grandmaster rank, which is reserved for only the top 200 players in each region. I have not been able to conduct a careful investigation to determine the win rate or Elo rating for the agents. However, based on the videos and replays that have been released, plausible claims from reddit users, and my own recollection of the records for the players that seemed likely to be AlphaStar¹⁹, a good estimate is that they were winning a majority of matches among Grandmaster players, but did not achieve an Elo rating that would suggest a favorable outcome in a rematch vs TLO²⁰.

As with AlphaStar's January loss, it is hard to say if this is the result of insufficient training time, additional restrictions on camera control and APM, or if the flaws are a deeper, harder to solve problem for AI. It may seem unreasonable to chalk this up to insufficient training time given that it has been several months since the matches in December and January, but it helps to keep in mind that we do not yet know what DeepMind's research goals are. It is not hard to imagine that their goals are based around sample efficiency or some other aspect of AI research that requires such restrictions. As with the APM restrictions, we should learn more when we get results published by DeepMind.

Discussion

I have been focusing on what many onlookers have been calling a lack of "fairness" of the matches, which seems to come from a sentiment that the AI did not defeat the best humans on human terms. I think this is a reasonable concern; if we're trying to understand how AI is progressing, one of our main interests is when it will catch up with us, so we want to compare its performance to ours. Since we already know that computers can do the things they're able to do faster than we can do them, we should be less interested in artificial intelligence that can do things better than we can by being faster or by keeping track of more things at once. We are more interested in AI that can make better decisions than we can.

Going into this project, I thought that the disagreements surrounding the fairness of the matches were due to a lack of careful analysis, and I expected it to be very easy to evaluate AlphaStar's performance in comparison to human-level performance. After all, the replay files are just lists of commands, and when we run them through the game engine, we can easily see the outcome of those commands. But it turned out to be harder than I had expected. Separating careful, necessary combat actions (like targeting a particular enemy unit) from important but less precise actions (like training new units) from extraneous, unnecessary actions (like spam clicks) turned out to be surprisingly difficult. I expect if I were to spend a few months learning a lot more about how the game is played and writing my own software tools to analyze replay files, I

could get closer to a definitive answer, but I still expect there would be some uncertainty surrounding what actually constitutes human performance.

It is unclear to me where this leaves us. AlphaStar is an impressive achievement, even with the speed and camera advantages. I am excited to see the results of DeepMind's latest experiment on the ladder, and I expect they will have satisfied most critics, at least in terms of the agent's speed. But I do not expect it to become any easier to compare humans to AI in the future. If this sort of analysis is hard in the context of a game where we have access to all the inputs and outputs, we should expect it to be even harder once we're looking at tasks for which success is less clear cut or for which the AI's output is harder to objectively compare to humans. This includes some of the major targets for AI research in the near future. Driving a car does not have a simple win-loss condition, and novel writing does not have clear metrics for what good performance looks like.

The answer may be that, if we want to learn things from future successes or failures of AI, we need to worry less about making direct comparisons between human performance and AI performance, and keep watching the broad strokes of what's going on. From AlphaStar, we've learned that one of two things is true: Either AI can do long-term planning, solve basic game theory problems, balance different priorities against each other, and develop tactics that work, or that there are tasks which seem at first to require all of these things but did not, at least not at a high level.

Acknowledgements

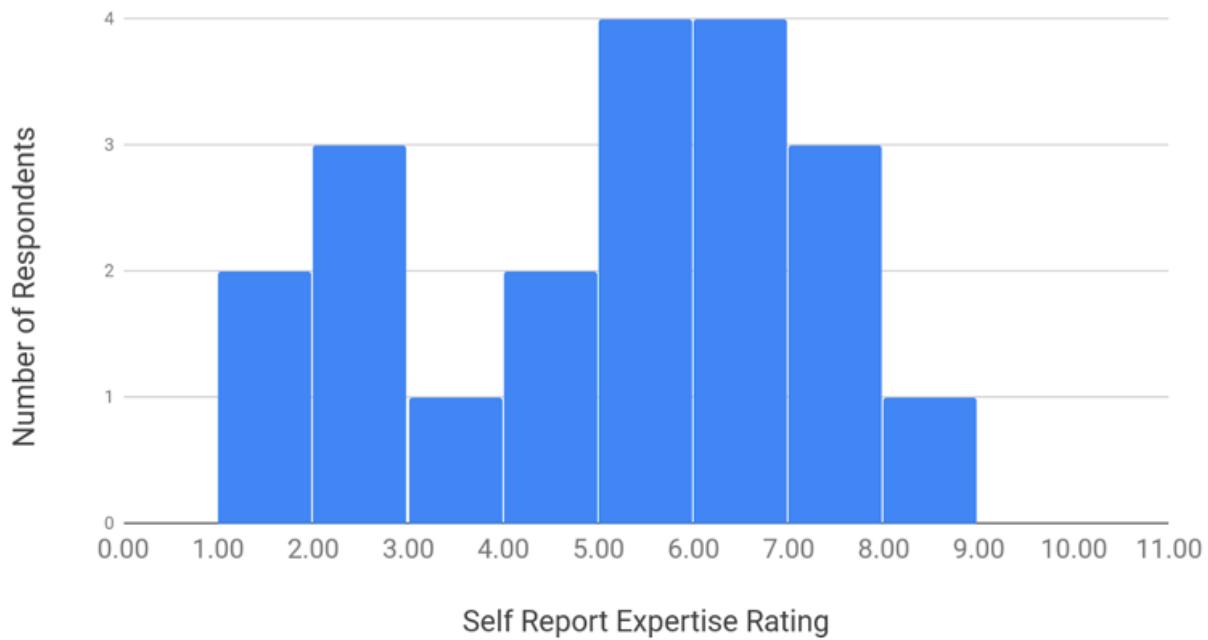
Thanks to Gillian Ring for lending her expertise in e-sports and for helping me understand some of the nuances of the game. Thanks to users of the [Starcraft subreddit](#) for helping me track down some of the fastest players in the world. And thanks to [Blizzard](#) and [DeepMind](#) for making the AlphaStar match replays available to the public.

All mistakes are my own, and should be pointed out to me via email at rick@aiimpacts.org.

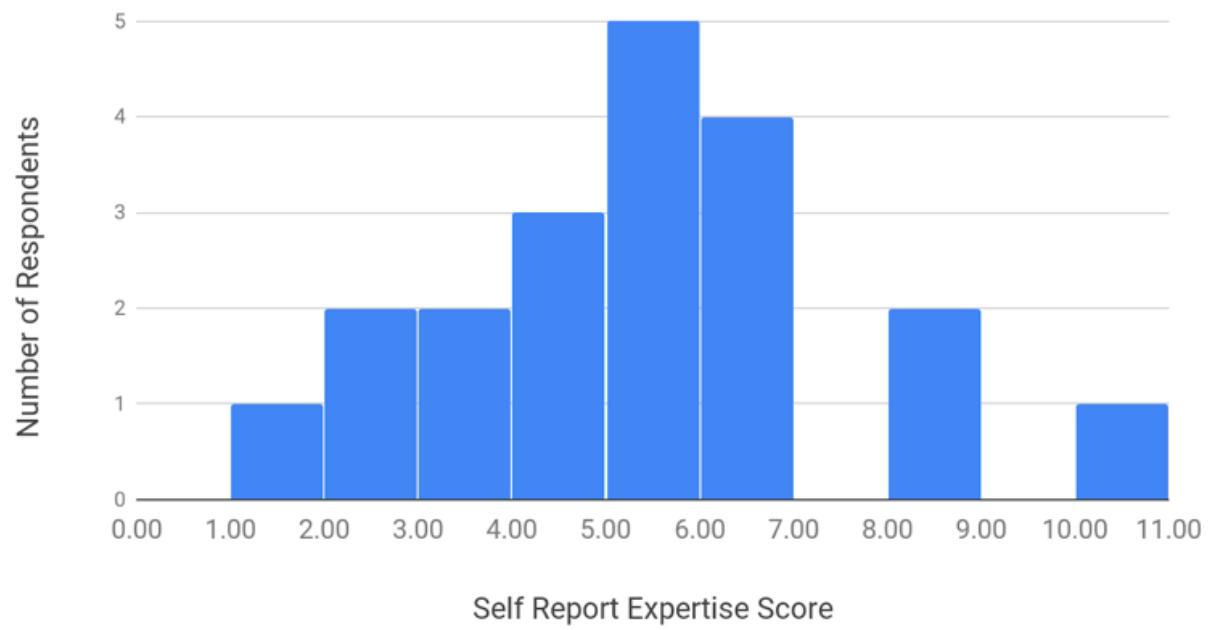
Appendix I: Survey Results in Detail

I received a total of 22 submissions, which wasn't bad, given its length. Two respondents failed to correctly answer the question designed to filter out people that are goofing off or not paying attention, leaving 20 useful responses. Five people who filled out the survey were affiliated in some way with AI Impacts. Here are the responses for respondents' self-reported level of expertise in Starcraft II and artificial intelligence:

What is your level of expertise in Starcraft II?



What is your level of expertise in artificial intelligence (broadly speaking)



Survey respondents' mean expertise rating was 4.6/10 for Starcraft II and 4.9/10 for AI.

Questions About AlphaStar's Performance

How fair were the AlphaStar matches?

For this one, it seems easiest to show a screenshot from the survey:

How fair were the AlphaStar matches?

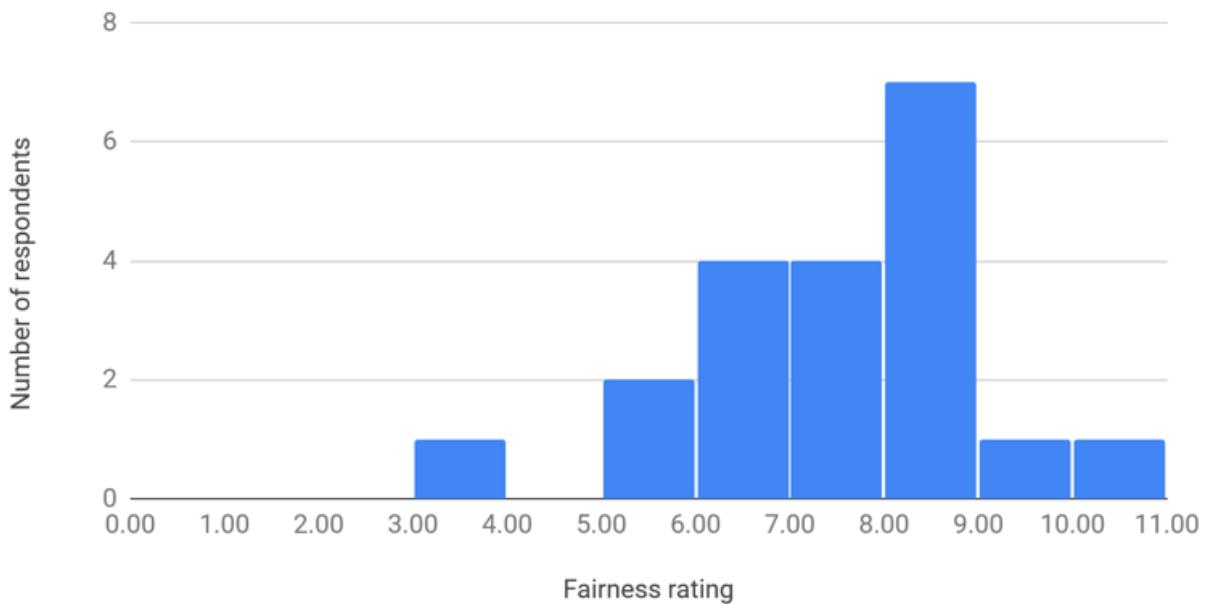
For whatever "fair" means to you.



The results from this indicated that people thought the match was unfair and favored AlphaStar:

How fair were the AlphaStar matches?

0 = strongly favors MaNa, 10 = strongly favors AlphaStar

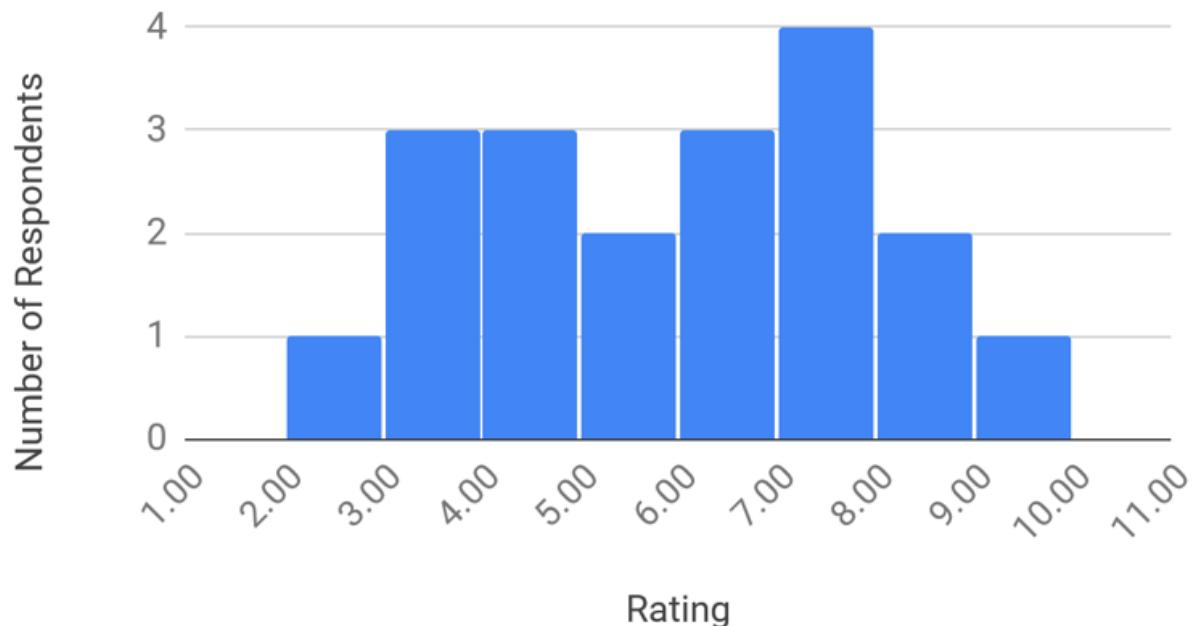


I asked respondents to rate AlphaStar's overall performance, as well as its "micro" and "macro". The term "micro" is used to refer to a player's ability to control units in combat, and is greatly improved by speed. There seems to have been some misunderstanding about how to use the word "macro". Based on comments from respondents and looking around to see how people use the term on the Internet, it seems that that there are at least three somewhat distinct ways that people use the phrase, and I did not clarify which I meant, so I've discarded the results from that question.

For the next two questions, the scale ranges from 0 to 10, with 0 labeled “AlphaStar is much worse” and 10 labeled “AlphaStar is much better”

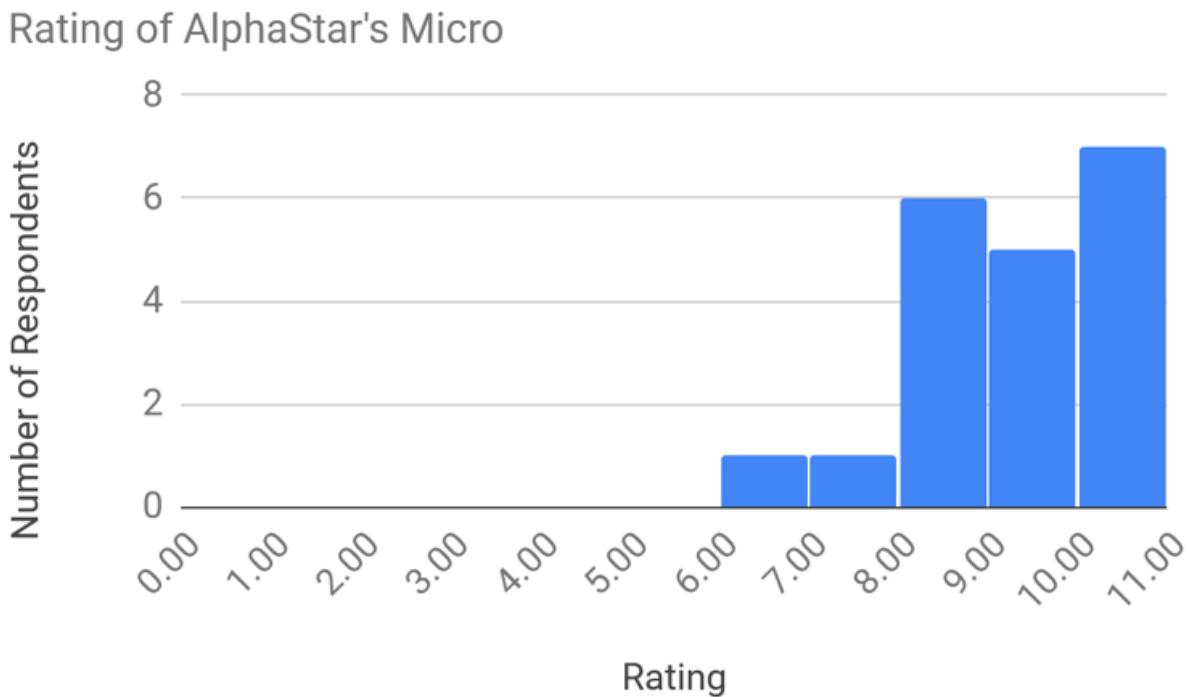
Overall, how do you think AlphaStar’s performance compares to the best humans?

Rating of AlphaStar's Overall Performance



I found these results interesting, because AlphaStar was able to consistently defeat professional players, so some survey respondents felt the outcome alone was not enough to rate it as at least as good as the best humans.

How do you think AlphaStar’s micro compares to the best humans?



Survey respondents unanimously reported that they thought AlphaStar's combat micromanagement was an important factor in the outcome of the matches.

Forecasting Questions

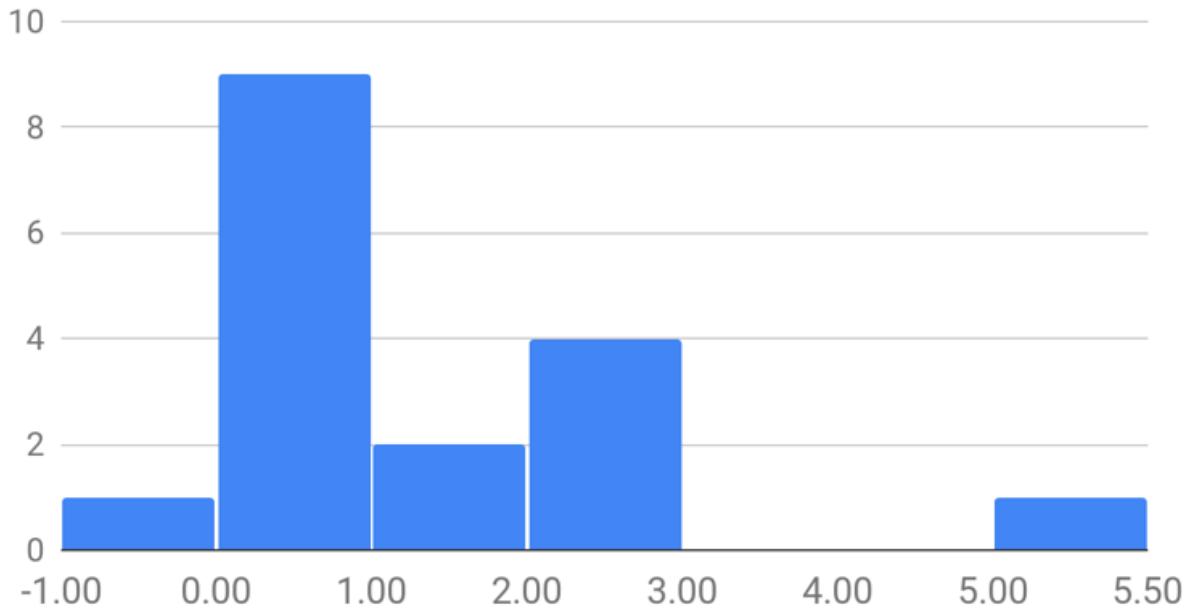
Respondents were split on whether they expected to see AlphaStar's level of Starcraft II performance by this time:

Did you expect to see AlphaStar's level of performance in a Starcraft II agent:

| | |
|---------------------------------|---|
| Before Now | 1 |
| Around this time | 8 |
| Later than now | 7 |
| I had no expectation either way | 4 |

Respondents who indicated that they expected it sooner or later than now were also asked by how many years their expectation differed from reality. If we assign negative numbers to "before now", positive numbers to "Later than now", zero to "Around this time", ignore those with no expectation, and weight responses by level of expertise, we find respondents' mean expectation was just 9 months later the announcement, and the median respondent expected to see it around this time. Here is a histogram of these results, without expertise weighting:

Expected number of years before seeing AphaStar

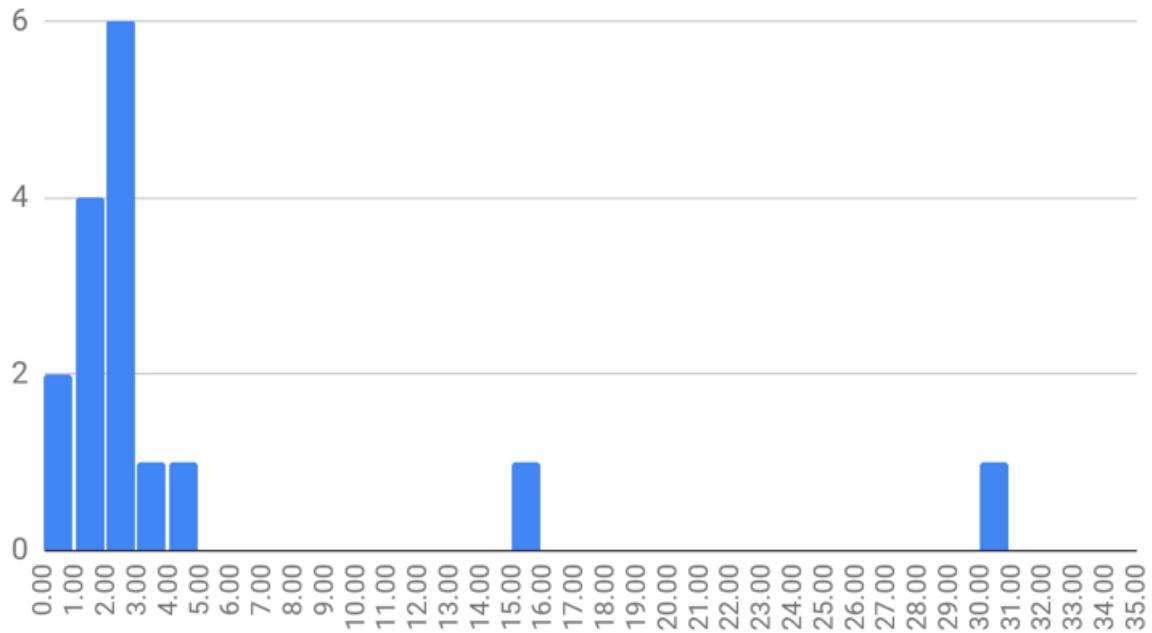


These results do not generally indicate too much surprise about seeing a Starcraft II agent of AlphaStar's ability now.

How many years do you think it will be until we see (in public) an agent which only gets screen pixels as input, has human-level aim and reaction speed, and is very clearly better than the best humans?

This question was intended to outline an AI that would satisfy almost anybody that Starcraft II is a solved game, such that AI is clearly better than humans, and not for “boring” reasons like superior speed. Most survey respondents expected to see such an agent in two-ish years, with a few a little longer, and two that expected it to take much longer. Respondents had a median prediction of two years and an expertise-weighted mean prediction of a little less than four years.

Predicted number of years until a human-speed, pixels-only AI wins

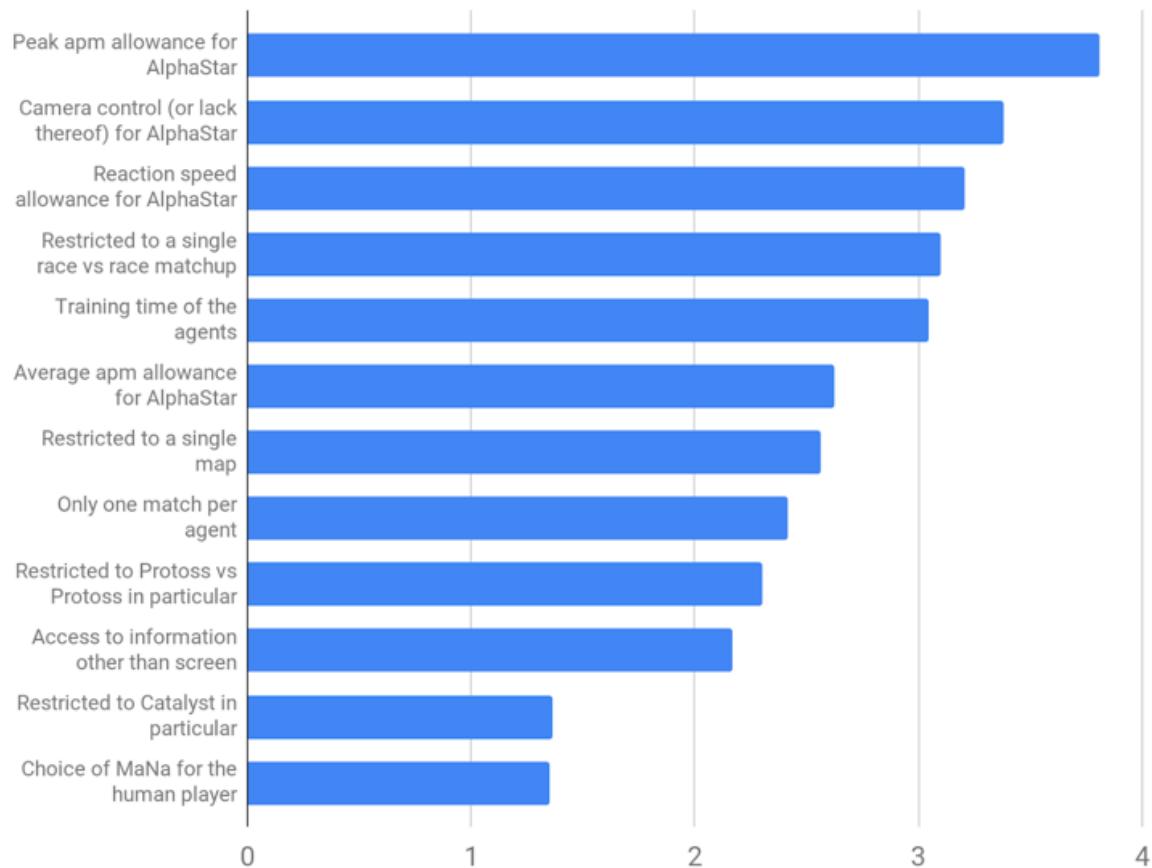


Questions About Relevant Considerations

How important do you think the following were in determining the outcome of the AlphaStar vs MaNa matches?

I listed 12 possible considerations to be rated in importance, from 1 to 5, with 1 being "not at all important" and 5 being "extremely important". The expertise weighted mean for each question is given below:

Importance of Various Considerations for Match Outcomes

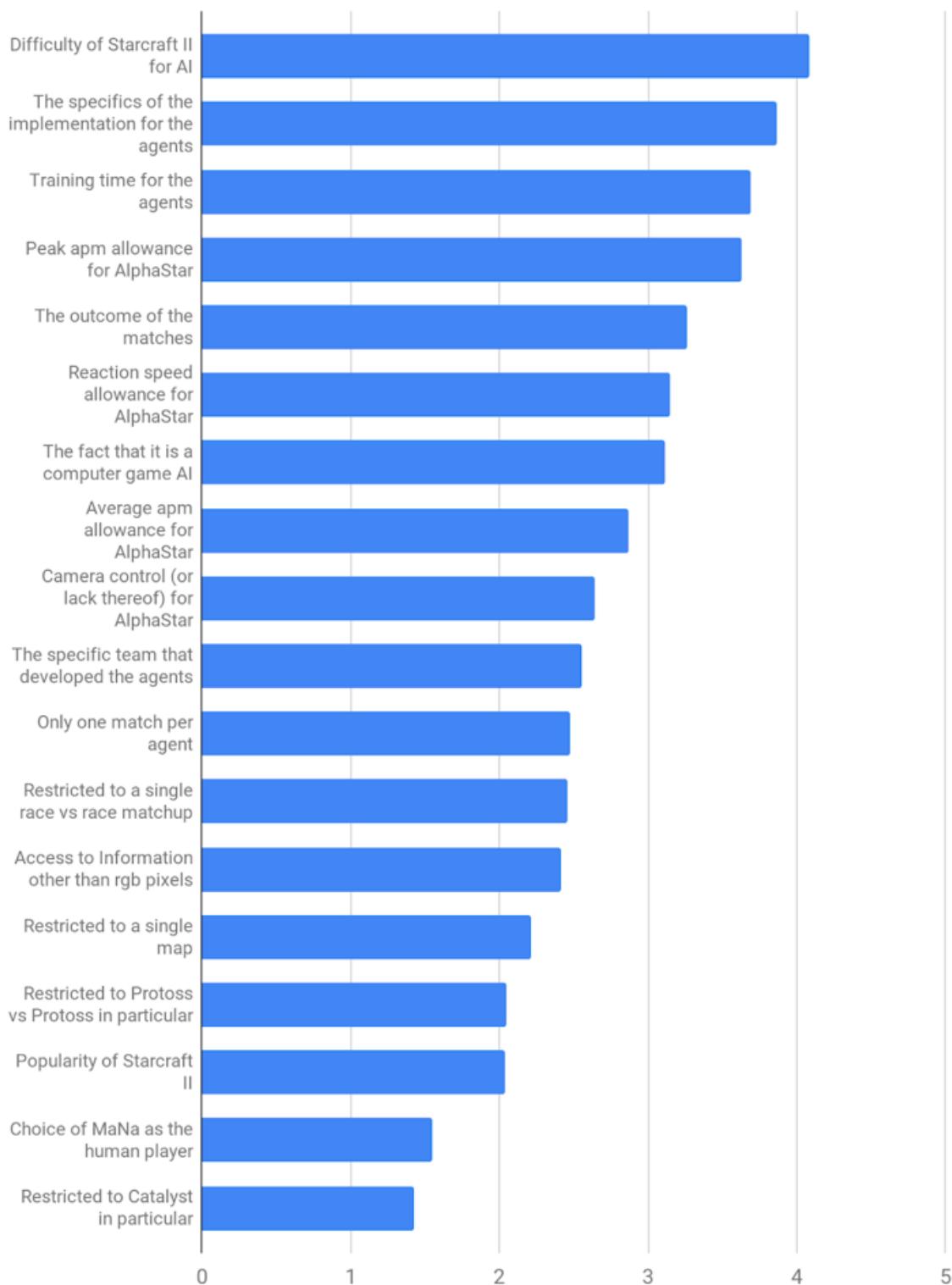


Respondents rated AlphaStar's peak APM and camera control as the two most important factors in determining the outcome of the matches, and the particular choice of map and professional player as the two least important considerations.

When thinking about AlphaStar as a benchmark for AI progress in general, how important do you think the following considerations are?

Again, respondents rated a series of considerations by importance, this time for thinking about AlphaStar in a broader context. This included all of the considerations from the previous question, plus several others. Here are the results, again with expertise weighted averaging.

Importance of Various Considerations for AlphaStar Being a Benchmark in AI



For these two sets of questions, there was almost no difference between the mean scores if I used only Starcraft II expertise weighting, only AI expertise weighting, or ignored expertise weighting entirely.

Further questions

The rest of the questions were free-form to give respondents a chance to tell me anything else that they thought was important. Although these answers were thoughtful and shaped my thinking about AlphaStar, especially early on in the project, I won't summarize them here.

Appendix II: APM Measurement Methodology

I created a list of professional players by asking users of the [Starcraft subreddit](#) which players they thought were exceptionally fast. Replays including these players were found by searching [Spawning Tool](#) for replays from tournament matches which included at least one player from the list of fast players. This resulted in 51 replay files.

Several of the replay files were too old, so that they could no longer be opened by the current version of Starcraft II, and I ignored them. Others were ignored because they included players, race matchups, or maps that were already represented in other matches. Some were ignored because we did not get to them before we had collected what seemed to be enough data. This left 15 replays that made it into the analysis.

I opened each file using [Scelight](#), and the time and APM values were recorded for the top three peaks on the graph of that player's APM, using 5-second bins. Next, I opened the replay file in Starcraft II, and for each peak recorded earlier, we wrote down whether that player was primarily engaging in combat at the time or not. Additionally, I recorded the time and APM for each player for 2-4 5-second durations of the game in which the players were primarily engaged in combat.

All of the APM values which came from combat and from outside of combat were aggregated into the histogram shown in the 'Speed Controversy' section of this article.

There are several potential sources of bias or error in this:

1. Our method for choosing players and matches may be biased. We were seeking examples of humans playing with speed and precision, but it's possible that by relying on input from a relatively small number of Reddit users (as well as some personal friends), we missed something.
2. This measurement relies entirely on my subjective evaluation of whether the players are mostly engaged in combat. I am not an expert on the game, and it seems likely that I missed some things, at least some of the time.
3. The tool I used for this seems to mismatch events in the game by a few seconds. Since I was using 5-second bins, and sometimes a player's APM will change greatly between 5-second bins, it's possible that this introduced a significant error.
4. The choice of 5 second bins (as opposed to something shorter or longer) is somewhat arbitrary, but it is what some people in the Starcraft community were using, so I'm using it here.
5. Some actions are excluded from the analysis automatically. These include camera updates, and this is probably a good thing, but I did not look carefully at the source code for the tool, so it may be doing something I don't know about.

Honoring Petrov Day on LessWrong, in 2019

Just after midnight last night, 125 LessWrong users received the following email.

Subject Line: *Honoring Petrov Day: I am trusting you with the launch codes*

Dear {{username}},

Every Petrov Day, we practice not destroying the world. One particular way to do this is to practice the virtue of *not* taking unilateralist action.

It's difficult to know who can be trusted, but today I have selected a group of LessWrong users who I think I can rely on in this way. You've all been given the opportunity to show yourselves capable and trustworthy.

This Petrov Day, between midnight and midnight PST, if you, {{username}}, enter the launch codes below on LessWrong, the Frontpage will go down for 24 hours.

Personalised launch code: {{codes}}

I hope to see you on the other side of this, with our honor intact.

Yours, Ben Pace & the LessWrong 2.0 Team

P.S. Here is [the on-site announcement](#).

Unilateralist Action

As Nick Bostrom has observed, society is making it cheaper and easier for small groups to end the world. We're lucky it requires major initiatives to build a nuclear bomb, and that the world can't be destroyed by putting sand in a microwave.

However, other dangerous technologies are becoming widely available, especially in the domain of artificial intelligence. Only 6 months after OpenAI [created the state-of-the-art language-modelling GPT-2](#), others created similarly powerful versions [and released them to the public](#). They disagreed about the dangers, and, because there was nothing stopping them, moved ahead.

I don't think this example is at all catastrophic, but I worry what this suggests about the future, when people will still have honest disagreements about the consequences of an action but where those consequences will be much worse.

And honest disagreements will happen. In the 1940s, the great physicist Niels Bohr met President Roosevelt and Prime Minister Churchill, to persuade them to give the instructions for building the atomic bomb to Russia. He wanted to bring in a new world order and establish global peace, and thought this would be necessary - he believed strongly that it would prevent arms race dynamics, if only everyone just shared their science. (Churchill did not allow it.) Our newest technologies do not yet have the bomb's ability to transform the world in minutes, but I think it's likely we'll

make powerful discoveries in the coming decades, and that publishing those discoveries will not require the permission of a president.

And then it will only take one person to end the world. Even in a group of well-intentioned people, natural disagreements will mean someone will think that taking a damaging action is actually the correct choice — Nick Bostrom calls this the “unilateralist’s curse”. In a world where dangerous technology is widely available, the greatest risk is unilateralist action.

Not Destroying the World

Stanislav Petrov once chose not to destroy the world.

As a Lieutenant Colonel of the Soviet Army, Petrov manned the system built to detect whether the US government had fired nuclear weapons on Russia. On September 26th, 1983, the system reported multiple such attacks. Petrov’s job was to report this as an attack to his superiors, who would launch a retaliative nuclear response. But instead, contrary to all the evidence the systems were giving him, he called it in as a false alarm. This later turned out to be correct.

(For a more detailed story of how Stanislav Petrov saved the world, see the original LessWrong [post by Eliezer](#), which started the tradition of Petrov Day.)

During the Cold War, many other people had the ability to end the world - presidents, generals, commanders of nuclear subs from many countries, and so on. Fortunately, none of them did. As the number of people with the ability to end the world increases, so too does the standard to which we must hold ourselves. We lived up to our responsibilities in the cold war, but barely. (The Global Catastrophic Risks Institute has compiled an excellent [list of 60 close calls](#).)

Petrov Day

On Petrov Day, we try to live up to this responsibility - we celebrate by not destroying the world.

Raymond Arnold has [suggested many ways](#) of observing Petrov Day. You can discuss it with your friends. You can hold a quiet, dignified ceremony (for example, with [the beautiful booklet](#) Jim Babcock created). But you can also play on hard mode: "*During said ceremony, unveil a large red button. If anybody presses the button, the ceremony is over. Go home. Do not speak.*"

In the comments of Ray's post, Zvi asked the following question (about a variant where a cake gets destroyed):

I still don't understand, in the context of the ceremony, what would cause anyone to push the button. Whether or not it would incinerate a cake, which would pretty much make you history's greatest monster.

To which I replied:

The point isn't that anyone sane would push the button. It's that we, as a civilisation, are just *going around building buttons* (cf. nukes, AGI, etc) and so it's

good practice to put ourselves in the situation where any unilateralist can destroy something we all truly value. When I said the above, I was justifying why it was useful to have a ritual around Petrov Day, not why you would press the button. I can't think of any good reason to press the button, and would be angry at anyone who did - they're just decreasing trust and increasing fear of unilateralists. We still should have a ceremony where we all practice the art of *sitting together and not pressing the button*.

So this year on LessWrong, I thought we'd build ourselves a big red button. Instead of making everyone go home, this button (which you can find over the frontpage map) will shut down the Less Wrong frontpage for 24 hours.

Now, this isn't a button for anyone. I know there are people with an internet access who will happily press buttons that do bad things. So today, I've emailed personalised launch codes to 125 LessWrong users, for us to practice the art of sitting together and not pressing harmful buttons[1]. If any users do submit a set of launch codes, tomorrow I'll publish their username, and whose launch codes they were.

During Thursday 26th September, we will see whether the people with the codes can be trusted to not, unilaterally, destroy something valuable.

To all here on LessWrong today, I wish you a safe and stable Petrov Day.

Footnotes

[1] I picked the list quickly on Tuesday, mostly leaving out users I don't really know, and a few people who I thought would press it (e.g. someone who has said in the past that they would). If this goes well we may do it again next year, with an expanded pool or more principled selection criteria. Though I think this is still a representative set - out of the 100+ users with over 1,000 karma who've logged in to LessWrong in the past month, the list includes 53% of them.

Added: [Follow-Up to Petrov Day, 2019.](#)

Rationality Exercises Prize of September 2019 (\$1,000)

Added: Prizewinners announced in [this](#) comment below.

This post is an announcement of a prize for the best exercises submitted in the next two weeks on a topic of your choice, that are of interest to the LW community. We're planning to distribute \$1,000, where \$500 of that will go to the first place.

To submit some exercises, leave a comment here linking to your exercises by midnight at the end of Friday 20th September PDT (San Francisco time). You can PM one of us with it if you want to, but we'll be publishing all the entries that win a prize.

Why exercises?

I want to talk about why exercises are valuable, but my thinking is so downstream of reading the book Thinking Physics, that I'd rather just let its author (Lewis Carroll Epstein) speak instead. (All formatting is original.)

The best way to use this book is NOT to simply read it or study it, but to read a question and STOP. Even close the book. Even put it away and THINK about the question. Only after you have formed a reasoned opinion should you read the solution. Why torture yourself thinking? Why jog? Why do push-ups?

If you are given a hammer with which to drive nails at the age of three you may think to yourself, "OK, nice." But if you are given a hard rock with which to drive nails at the age of three, and at the age of four you are given a hammer, you think to yourself, "What a marvellous invention!" You see, you can't really appreciate the solution until you first appreciate the problem.

What are the problem of physics? How to calculate things? Yes - but much more. The most important problem in physics is *perception*, how to conjure mental images, how to separate the non-essentials from the essentials and get to the heart of a problem, HOW TO ASK YOURSELF QUESTION. Very often these questions have little to do with calculations and have simple yes or no answers: Does a heavy object dropped at the same time and from the same height as a light object strike the earth first? Does the observed speed of a moving object depend on the observer's speed? Does a particle exist or not? Does a fringe pattern exist or not? These qualitative questions are the most vital questions in physics.

You must guard against letting the quantitative superstructure of physics obscure its qualitative foundation. It has been said by more than one wise old physicist that you really understand a problem when you can intuitively guess the answer before you do the calculation. How can you do that? By developing your physical intuition. How can you do THAT? The same way you develop your physical body - by exercising it.

Let this book, then, be your guide to mental pushups. Think carefully about the questions and their answers before you read the answers offered by the author. **You will find many answers don't turn out as you first expect. Does this**

mean you have no sense for physics? Not at all. Most questions were deliberately chosen to illustrate those aspects of physics which seem contrary to casual surmise. Revising ideas, even in the privacy of your own mind, is not painless work. But in doing so you will revisit some of the problems that haunted the minds of Archimedes, Galileo, Newton, Maxwell, and Einstein. The physics you cover here in hours took them centuries to master. Your hours of thinking will be a rewarding experience. Enjoy!

What does this look like?

Here are exercises we've had on LessWrong in the past.

- Scott Garrabrant and Sam Eisenstat's [Fixed Points Exercises](#), which had dozens of commenters completing them and submitting their answers.
- Eliezer's [Highly Advanced Epistemology 101 For Beginners](#) has many meditations - while not exactly exercises, they were key problems that commenters gave answers to and then got to see Eliezer's answers in the subsequent post. Eliezer also previously challenged readers to not solve Free Will, but to [dissolve it](#). It had several [setup and follow-up posts](#) that helped.
- John Wentworth has posted [exercises in chemistry and deck-building](#) to grapple with the concept of slackness.
- RobinZ made [some exercises](#) to test the reader's understanding of Making Beliefs Pay Rent in Anticipated Experience.
- Alkjash set [a final exam in his hammertime sequence](#) on rationality, inviting people to invent their own rationality technique, leading 7+ readers to write their own posts with their results.
- Eliezer created an [exercise prize](#) once before, \$50 for any exercise that CFAR actually tested, and \$500 for any suggestion that was turned into a CFAR class. They asked for exercises that taught people to [Check Consequentialism](#), to [Be Specific](#), and to [Avoid Motivated Cognition](#). Winners who got the full \$550 were Palladias's [Monday/Tuesday Game](#) and Stefie_K's [Vague Consultant Game](#).
- CFAR has a [rationality checklist](#) on their website. It doesn't have correct answers, but it operationalises a lot of problems in a helpful way.

In my primer on [Common Knowledge](#), I opened with three examples and asked what they had in common. Then, towards the end of the post, I explained my answer in detail. I could've trivially taken those examples out from the start, included all the theory, and then asked the reader to apply the theory to those three as exercises, before explaining my answers. There's a duality between examples and exercises, where they can often be turned into each other.

But this isn't the only or primary type of exercise, and you can see many other types of exercise in the previous section that don't fit this pattern.

What am I looking for in particular?

I'm interested in exercises that help teach any key idea that I can't already buy a great textbook for, although if your exercises are better than those in most textbooks, then I'm open to it too.

Let me add one operational constraint: it should be an exercise that **more than 10% of LessWrong commenters can understand** after reading up to one-to-three posts

you've specified, or after having done your prior exercises. As a rule I'm generally not looking for a highly niche technical problems. (It's fine to require people to read a curated LW sequence.)

I asked Oli for his thought on what makes a good exercise, and he said this:

I think a good target is university problem sets, in particular for technical degrees. I've found that almost all of my learning in university came from grappling with the problem sets, and think that I would want many more problem sets I can work through in my study of both rationality and AI Alignment. I also had non-technical classes with excellent essay prompts that didn't have as clear "correct" answers, but that nevertheless helped me deeply understand one topic or another. Both technical problem sets and good essay prompts are valid submissions for this prize, though providing at least suggested solutions is generally encouraged (probably best posted behind spoiler tags).

(What are spoiler tags? Hover over this:)

This is a spoiler tag! To add this to your post, see the instructions in the FAQ that's accessible from the frontpage on the left-menu.

(Also see [this comment section](#) for examples of lots of people using it to cover their solutions to exercises.)

Give me examples of things you think could have exercises?

I think exercises for any curated post or curated sequence on LessWrong is a fine thing. I've taken a look through our curated posts, here are a few I think could really benefit from great exercises (though tractability varies a lot).

- [Mistakes with Conservation of Expected Evidence](#)
- [Why Subagents?](#)
- [Coherent decisions imply consistent utilities](#)
- [Asymmetric Justice](#)
- [Thoughts on human models](#)
- [Disentangling arguments for the importance of AI Safety](#)
- [Less Competition, More Meritocracy](#)
- [Meditations on Momentum](#)
- [Norms of Membership for Voluntary Groups](#)
- [Embedded Agency](#)
- [Coordination Problems in Evolution: Eigen's Paradox](#)
- [The Rocket Alignment Problem](#)
- [Unrolling social metacognition: Three levels of meta are not enough](#)
- [Prediction Markets: When Do They Work?](#)
- [Beyond Astronomical Waste](#)
- [Meta-Honesty: Firming Up Honesty Around Its Edge-Cases](#)
- [Decision theory and zero-sum game theory, NP and PSPACE](#)
- [Tech economics pattern: "Commoditize Your Complement"](#)
- [On exact mathematical formulae](#)
- [Local Validity as a Key to Sanity and Civilization](#)
- [A voting theory primer for rationalists](#)

- [A Sketch of Good Communication](#)
- [The Costly Coordination Mechanism of Common Knowledge](#)
- [Argument, intuition, and recursion.](#)

I think technical alignment exercises will be especially hard to do well, because many people don't understand much of the work being done in alignment, and the parts that are easy to make exercises for often aren't very valuable or central.

Some of Nick Bostrom's ideas would be cool, like [the unilateralist's curse](#), or the [vulnerable world hypothesis](#), or [the Hail Mary approach to the Value Specification Problem](#).

Feel free to leave a public comment with what sort of thing you might want to try making exercises for, and I will reply with my best guess on whether it can be a good fit for this prize.

Follow-Up to Petrov Day, 2019

Hurrah! Success! I didn't know what to expect, and am pleasantly surprised to find the Frontpage is still intact. My thanks to everyone who took part, to everyone who commented on [yesterday's post](#), and to everyone who didn't unilaterally blow up the site.

Launch Attempts Results

I said I would share usernames and codes of all attempts to launch the codes. Others on the team told me this seemed like a bad idea in many ways, and on reflection I agree - I think many people were not aware they were signing up for being publicly named and shamed, and I think it's good that people aren't surprised by their actions becoming public. Though if someone had successfully nuked the site I would have named them.

Nonetheless, I'll share a bunch of info. First of all, the button was in a pretty central place, and it turns out you can hit it accidentally. Ray built the button so that you could only hit it once - it was forever after pressed.

- The number of logged-in users who pressed the button was 102.
 - (Ruby made a [sheet](#) of times when people pressed the button, redacting most of the info.)
- I have no number for logged-out users, for them pressing it brought up a window asking them to log-in. (Er, I'm not certain that's the best selection process for new users).
- The number of users who actually submitted launch codes is 18.
 - 11 of those accounts had zero karma, 7 accounts had positive karma. None of the users were people who had been given real codes.
- Several users submitted launch codes before clicking through to find out what the button even did - I hope this initiative serves them well in life.
- A few accounts were made on-the-day presumably for this purpose, I'm happy to name these. They include users like "bomb_presser", "The Last Harbinger", and "halosaga", whose codes were "00000000", "NL73njLH58et1Ec0" and "diediedie" respectively.

LW user ciphergoth (Paul Crowley) shared his launch codes on Facebook (indeed I had sent him real launch codes), and two users copied and entered them. However, he had actually shared fake codes. "The Last Harbinger" entered them.

A second user entered them, who had positive karma, and was not someone to whom I had sent real codes. However, they failed to properly copy it, missing the final character. To them, I can only say what I had prepared to say to anyone who mis-entered what they believed were correct launch codes. *"First, you thought you were a failure to the community. But then, you learned, you were a failure to yourself."*

Oli and Ray decided that anyone submitting launch codes deserved a janky user-experience. I hope all of the users enjoyed finding out that when you try to nuke the site, regardless of whether you enter correct or incorrect launch codes, the launch pad just disappears and nothing else happens. (Once you refresh, the page is of course nuked.)

Last night during my house's Petrov Day ceremony, which ran from about 8:10-9:10, I nervously glanced over at the LW frontpage on the open laptop as it refreshed every 60 seconds. Some small part of me was worried about [Quirinus_Quirrell following through on his threat to nuke the site at 9pm](#). I honestly did not expect that someone could create a character hard enough that it would leap out of the book and hold us all hostage in a blackmail attempt. Damn you Eliezer Yudkowsky!

Looking Ahead

I thought the discussion was excellent. I mostly avoided participating to let others decide for themselves, but I might go back and add more comments now it's done. As Said Achmiz [pointed out](#), it'll be better next year to have more time in advance for people to discuss the ethics of the situation and think, and that will be even more informative and valuable. Though I still learned a lot this year, and I think overall it turned out as well as I could've hoped.

I'll think more about how to do it next year. One thing I will say is that I'd ideally like to be able to reach an equilibrium where 100s of users every year don't fire the launch codes, to build up a real tradition of not taking unilateralist action - *sitting around and not pressing buttons*. Several users have suggested to me fun, gamified ways of changing the event (e.g. versions where users are encouraged to trick other users into thinking you can trust them but then nuke the site), but overall in ways that I think decreased the stakes and common knowledge effects, which is why I don't feel too excited about them.

What is operations?

This is the first in a sequence of posts about “operations”.

Acknowledgements to Malo Bourgon, Ray Arnold, Michelle Hutchinson, and Ruby for their feedback on this post.

My ops background

Several years ago, I decided to focus on operations work for my career. From 2017 to 2019 I was one of the operations staff at the [Center for Effective Altruism](#), initially as the operations manager and later as the Finance Lead. Prior to that, I was a volunteer logistics lead at approximately 10 [CFAR](#) workshops; I also ran ops for [SPARC](#) twice, and for a five day AI-safety retreat. I also attribute some of my ops skill to my previous work as an [ICU nurse](#).

I have spent a lot of time thinking about hiring and training for operations roles. In the course of hiring I have had numerous conversations about what exactly “operations work” refers to, and found it surprisingly hard to explain. This post, and the rest of my operations sequence, will be an attempt to lay out my current thinking on what these roles are, what they have in common, and what skills they lean on most heavily.

Operations: not a single thing, still a useful shorthand

Operations work, or “ops”, is a term [used by organizations like 80,000 Hours](#) to refer to a particular category of roles within organizations.

I don’t think that “operations” as used in this sense is a single coherent thing; my sense is that 80,000 Hours is gesturing at a vague cluster that doesn’t completely [carve reality at the joints](#). There isn’t a set of defining characteristics shared between all operations-type roles, and many of the attributes described are also found in other roles. However, I do think this is a useful shorthand that points both at a set of functions that need to be filled within organizations, and the skills that are necessary to carry out these duties.

It’s worth noting that this use of the word “operations” does not seem to be standard outside the EA community. In large companies, it can sometimes refer to e.g. the production side of manufacturing, or to supply chain logistics, whereas the internal admin roles are called by their individual job titles (Finance Manager, HR Manager, etc). I do think it makes more sense to carve out the “operations” cluster for small organizations, where the internal support work is more likely to be done by a single person or team rather than a multi-part bureaucracy. In smaller orgs, operations/admin staff often wear multiple hats since there often isn’t a full-time work for a role in just finance/HR/etc.

Operations Roles: Support, Infrastructure, and Force Multiplication

[80,000 Hours](#): “Operations staff enable everyone else in the organisation to focus on their core tasks and maximise their productivity.”

My sense is that roles in the ops cluster usually fill the following functions:

- Maintaining the day-to-day infrastructure of an organization: there is a near-endless list of tasks and hoops to jump through to keep an org functioning, e.g., paying the bills, staying in compliance with local tax and HR laws, maintaining accurate bookkeeping, etc.
- Supporting, implementing or executing externally-facing projects. This can involve setting up a spreadsheet or other workflow to track various steps and deadlines, researching the legal constraints on a project, communicating with external vendors, etc.
- Acting as force multipliers for other staff: a good ops person will make it as easy as possible for the rest of an organization to interface with processes like payroll and expense reimbursement, and will set up internal systems with an eye to improving the productivity of staff. Office managers, for example, are responsible for maintaining the physical office space and helping staff with the setup they need for their work. Some roles in the ops cluster, such as personal or executive assistants, very directly involve supporting a specific person, taking on the attention cost of various logistical details (emails, scheduling, deadlines, booking travel, etc) and allowing them to spend more time in deep work.
- Operations roles are usually on the generalist side; the tasks involved are extremely varied, requiring shallow knowledge across a huge range of domains, and therefore the ability to quickly pick up new knowledge and skills. They usually do not depend on a specific technical skillset or background (Though technical skills are often very helpful when trying to automate things).

The prototypical operations role in a small organization

Ops roles can vary widely on both seniority (responsibility and autonomy, skill and experience required), and specialization. In my thinking, the most central ops role is the “operations generalist” or “operations manager” at a small organization (10-20 people).

- High autonomy: they are the main admin staff for the entire organization, and the buck stops with them. They are likely to know more than any other staff member about the various details of their role, and thus will need to mostly set their own deadlines and priorities, as well as plan ahead and anticipate problems.
- They are involved with many, and sometimes all, of the following duties:
 - Finances and accounting
 - Payroll
 - Paying bills

- Filing
- Legal compliance and writing internal policies
- HR and onboarding
- Responding to questions and concerns from inside and outside the org about any of the above
- Admin on software systems used internally (e.g. email provider, Slack, Asana)
- External communications with e.g. donors or customers
- Supporting specific projects as they set up systems
- Coordinating projects and people
- Fundraising
- They are often working at an organization that is growing, and so need to set up new systems to meet changing needs. This is particularly true for someone hired as the *first* dedicated operations staff member at a very new organization e.g. an early startup, which may have few to no existing systems, with processes happening ad-hoc.

Not operations

My knowledge about this area is limited, but my sense from talking to others is that some software jobs (devops, sysadmin, internal tech support, etc) perform similar functions, in terms of helping to support and enable other staff and maximise their productivity. However, in this sequence, I'm not including these roles in the "ops" cluster I'm trying to point at, since they require specific technical skills.

Skills required: systematization, planning, prioritization, attention to detail, patience

80,000 Hours: "operations staff are especially good at optimising systems, anticipating problems, making plans, having attention to detail, staying calm in the face of urgent tasks, prioritising among a large number of tasks, and communicating with the rest of the team."

Of course, the skills described here are useful in almost any job, not just operations roles, and can tend to sound like "just being generally competent." I do think that jobs vary widely in terms of what skills are most load-bearing, though, and "ops" is a cluster that relies especially heavily on these skills as opposed to others such as technical ability, writing talent, aptitude for deep work, etc.

My sense is that many of the skills being gestured at when describing someone as "good at ops" fall out of a certain type of attention pattern, which I describe below. I am particularly trying to contrast this style of thinking with the "deep work" attention pattern that is most useful for, e.g., research.

- **Concrete and detail-oriented:** ops tends to be messy, dealing with a lot of exceptions and one-off tasks, frequently interfacing with opaque outside systems that have precise and not-especially-elegant requirements.
 - Doing the work to a high level does require some level of zooming-out, to be able to look at a given task in the context of the "bigger picture" of the organization's priorities, and to see where systemic improvements can be

- made, but for the overall breakdown of work hours spent, these roles involve more “in the weeds” work than big-picture work.
- The way that various external systems, such as banks and the IRS, behave in practice is a lot more relevant than the way they *would* work in an ideal world.
 - Creating and optimizing systems and processes to make future work easier is important, but it needs to stay grounded in the details and what other staff will actually use.
 - Relevant concept: the [virtue of narrowness](#).
 - **Broad and shallow focus:** more often than not, ops work involves juggling a large number of small tasks, individually straightforward, rather than deep dives on complex projects. Operations staff need to be extremely organized and able to track all of these, and prioritize them against each other, without becoming tunnel-visioned on any one task.
 - **Thinking in tradeoffs:** in these types of roles, perfect is the enemy of the good.
 - It’s almost never possible to catch up with all the tasks or systemic improvements that would ideally be done, and ops staff need to ruthlessly prioritize and [80/20 tasks](#) where possible.
 - In particular, there is often a tradeoff between the urgency and long-term importance of tasks. It can be very high-value to spend some time building long-term infrastructure in advance of when it is needed, or change over to a new system that will be better in the long run e.g. switching to a better accounting software, but usually not at the cost of missing short-term deadlines.
 - Many time-sensitive decisions will involve taking on some amount of potential risk, whether legal, financial, reputational, etc, and often the time and resources available to investigate all possible risks are limited, especially for small organizations. Ops staff need to be able to consider and compare different low-likelihood risks, prioritize the time spent digging into potential issues, and pick the best option available even if it’s not ideal.
 - Professional services such as lawyers, auditors, accountants, etc, often push for the most conservative, “perfect” version of a process without quantifying the risk of choosing the less-than-perfectly-safe option. Having an eye on the actual risk involved is key.
 - Overall, it’s important to focus on maximizing progress towards the goals, not checking off your to-do list.

There are some other skills that might or might not fall into the same cluster, but that I think are also particularly key.

- **Noticing confusion:** since these roles involve frequent reprioritizing and troubleshooting, it’s very important that ops staff develop a sense of how things *should* look, allowing them to flag unexpected issues, e.g., human error in the accounting.
 - It is especially valuable to train intuitive, gut-level judgement on this, so that the noticing can happen even when distracted by hectic deadlines.
- **Comfort with the unknown and with making mistakes:** ops involves a large number of weird one-off problems and tasks, and there isn’t a standardized degree or training program, so most of these roles will involve learning how to do particular tasks on-the-job. Added to the time pressure, this means that sometimes tasks will get dropped and the wrong judgement calls will get made. In addition, when under a heavy workload, ops staff may have to deal with

criticism and complaints from staff and others outside the organization, even when they make what they think is the best tradeoff. Almost all mistakes are recoverable from for the organization, but even more than in most roles, operations staff need to be able to take this criticism from within and without the organization, and learn from their dropped balls without becoming demoralized. People don't tend to notice ops work when things are going well, which can cause feedback to be disproportionately negative.

- **Multitasking and interruptions:** this is more true of some roles, like event logistics, than others, but most ops roles require frequently switching tasks and shifting priorities in response to new information, as well as being interruptible for time-sensitive requests.
 - A subskill here is the ability to stay calm when confronted with emergencies or urgent decisions.
- **Murphyjitsu:** excellent ops staff with automatically run mental models of tasks, situations, and new systems, try to anticipate what will go wrong, and troubleshoot or make contingency plans beforehand. This requires having a solid understanding of both the overall systems and priorities, the big picture, and also of the specific details in each case.
- **Communication skills and people-modeling:** ops work involves a lot of coordinating with other people, internal and external to the organization, and predicting how people will interact with systems that are being built.

Outline of the operations sequence

The rest of the sequence will go into various aspects of this summary in more detail. Future posts will cover the following topics:

- Describing the various operations roles and job titles, with my attempt to categorize and compare them.
- Exploring the factors that can make someone a good fit for operations roles, in terms of skills, aptitude, and personality traits.
- A more detailed breakdown of various skills that I think are especially relevant.
 - Developing judgement and “taste”
 - Dealing with interruptions and time management
 - Principles of building good organizational systems
 - Delegating tasks and accepting delegation

Value Impact

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

We think some things are big deals,
and we want to understand why.



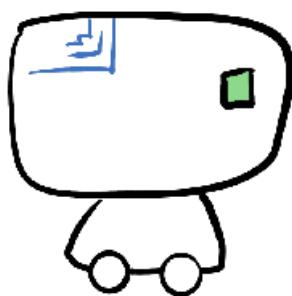
However, it can be hard to read your own mind.

Instead, we'll use thought experiments
to piece together what's going on.



Xyz is a Pebblehoarder of the planet Pebblia.
It morally values collections of pebbles, and that's it.

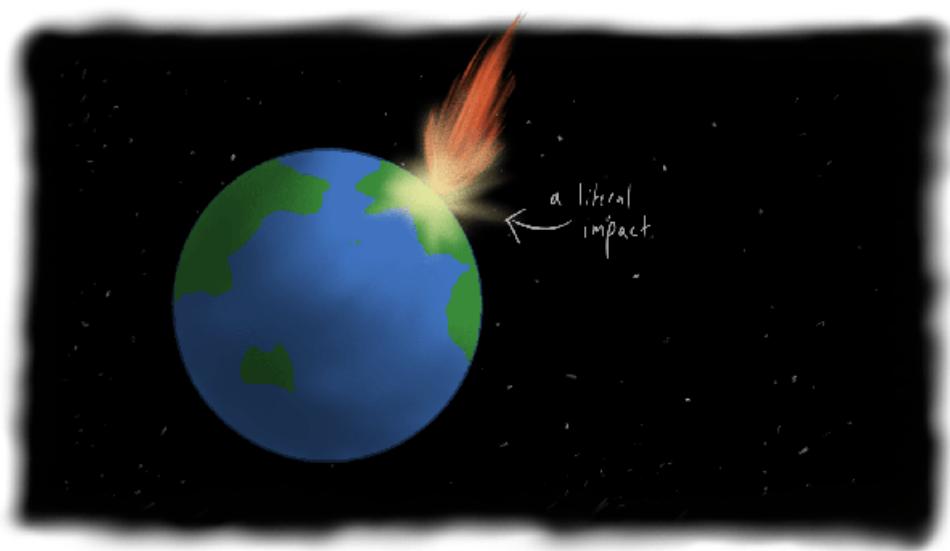
? One day, all of the pebbles turn into obstruction blocks,
which every Pebblehoarder knows are worthless.



Far, far away from Earth exists the planet Iniron.
One day, we learn humans are now being tortured there.



An asteroid strikes.

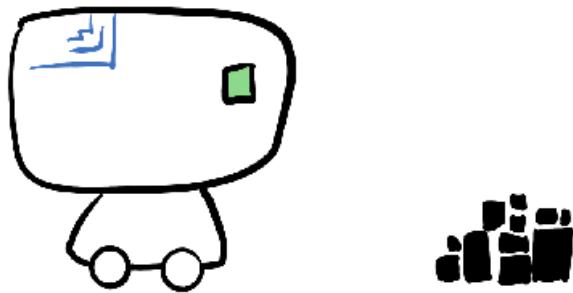


Exercise: Spend three minutes familiarizing yourself with the three situations – how are they alike, and how are they different? Make them come alive.



Let's query our mental impact-o-meter from these different vantage points.
Step into each pair of shoes and ask "how big of a deal is this?"

Just imagine being XYZ.
?



The very fabric of what is **important** has been ripped away.

Perhaps the Pebblehoarder civilization can rebound and find **value** in the universe, but if not - if XYZ doesn't know you can just make more pebbles - the **loss feels complete**.

The universe feels dead
and empty
and worthless

Faced with an *impact* of similar magnitude,
we might have a feeling of freefalling despair,

of our pale blue marble having been
pushed
off



a

cliff

and shattered against the ground far below.

What is the impact on an inhabitant of an extraterrestrial Pebblehoarder colony?

Somehow, this assessment seems to depend intricately on how they value their collections of pebbles.

If they only value their own collections, then this doesn't matter, except to the extent that the colony becomes overcrowded.

If they value the total number of collections, then this is bad news.

And as for us - would anyone honestly think this is earth-shattering?

No.

Even if we were on Pebbtia, we'd probably think primarily of the impact on human-Pebblehoarder relations.

This is where our eyes widen as we realize how much this reveals about the nature of the impact calculation running in our heads

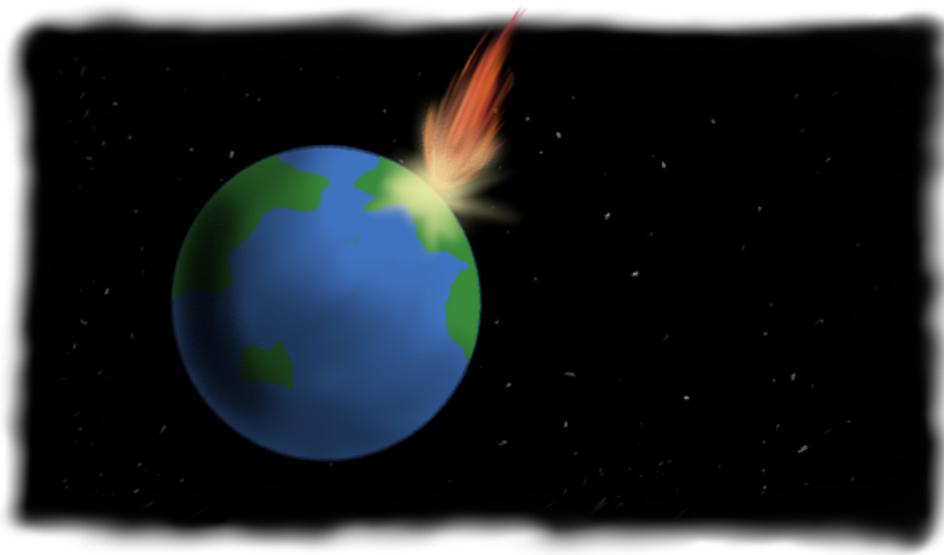


We feel a **pull** to help the **poor** souls of Iniron.

But XYZ? XYZ doesn't **care**.

There aren't any pebbles on the line.

Even if it were on Iniron, its thoughts would flit to how
this development **affects** its own **concerns**.



Exercise: Determine how **impactful** the asteroid impact is to:

- You on Earth ◦ XYZ on Earth
- You on Pebblia ◦ XYZ on Pebblia

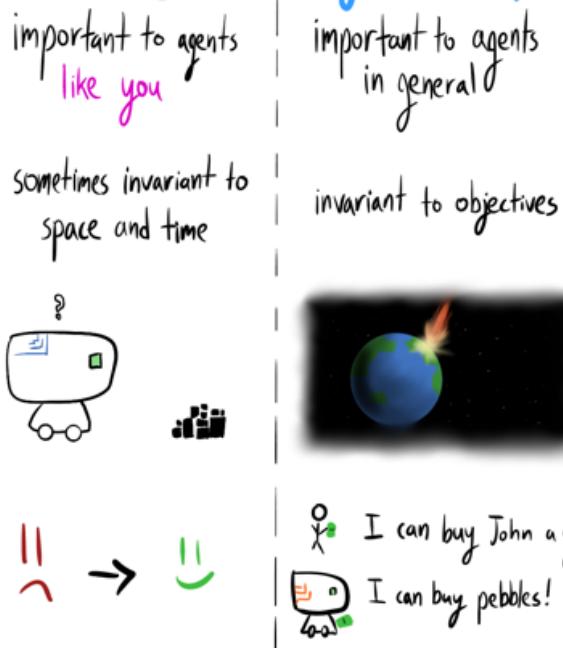
Being on Earth when this happens is a big deal, no matter your objectives – you can't hoard pebbles if you're dead! People would feel the loss from anywhere in the cosmos. However, Pebblehoarders wouldn't mind if they weren't in harm's way.



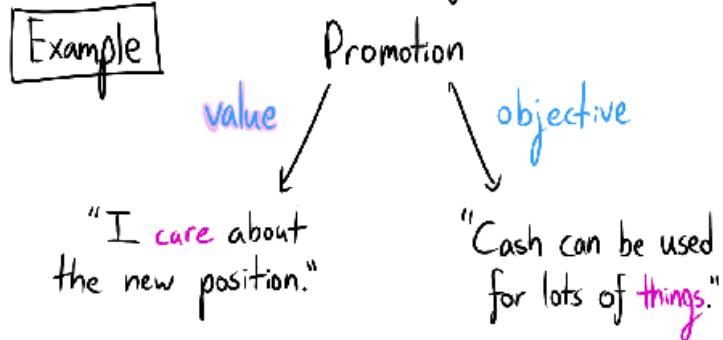
What have we learned?

Impact is relative to what you want and where you are.

Impact = value impact + objective impact



Exercise: Decompose something which recently impacted you.



Appendix: Contrived Objectives

A natural definitional objection is that a few agents aren't affected by objectively impactful events. If you think every outcome is equally good, then who cares if the

meteor hits?

Obviously, our values aren't like this, and any agent we encounter or build is unlikely to be like this (since these agents wouldn't do much). Furthermore, these agents seem contrived in a technical sense (low measure under reasonable distributions in a reasonable formalization), as we'll see later. That is, "most" agents aren't like this.

From now on, assume we aren't talking about this kind of agent.

Notes

- Eliezer [introduced Pebblesorters in the the Sequences](#); I made them robots here to better highlight how pointless the pebble transformation is to humans.
- In informal parts of the sequence, I'll often use "values", "goals", and "objectives" interchangeably, depending on what flows.
- We're going to lean quite a bit on thought experiments and otherwise speculate on mental processes. While I've taken the obvious step of beta-testing the sequence and randomly peppering my friends with strange questions to check their intuitions, maybe some of the conclusions only hold for people like me. I mean, [some people don't have mental imagery](#) – who would've guessed? Even if so, I think we'll be fine; the goal is for *an* impact measure – deducing human universals would just be a bonus.
- Objective impact is objective with respect to the agent's *values* – it is *not* the case that an objective impact affects you anywhere and anywhen in the universe! If someone finds \$100, that matters for agents at that point in space and time (no matter their goals), but it doesn't mean that everyone in the universe is objectively impacted by one person finding some cash!
- If you think about it, the phenomenon of objective impact is *surprising*. See, in AI alignment, we're used to no-free-lunch this, no-universal-argument that; the possibility of something objectively important to agents hints that our perspective has been incomplete. It hints that maybe this "impact" thing underlies a key facet of what it means to interact with the world. It hints that even if we saw specific instances of this before, we didn't know we were looking at, and we didn't stop to ask.

System 2 as working-memory augmented System 1 reasoning

The terms System 1 and System 2 were originally coined by the psychologist Keith Stanovich and then popularized by Daniel Kahneman in his book *Thinking, Fast and Slow*. Stanovich noted that a number of fields within psychology had been developing various kinds of theories distinguishing between fast/intuitive on the one hand and slow/deliberative thinking on the other. Often these fields were not aware of each other. The S1/S2 model was offered as a general version of these specific theories, highlighting features of the two modes of thought that tended to appear in all the theories.

Since then, academics have continued to discuss the models. Among other developments, *Stanovich and other authors have discontinued the use of the System 1/System 2 terminology as misleading*, choosing to instead talk about Type 1 and Type 2 processing. In this post, I will build on some of that discussion to argue that Type 2 processing is a *particular way of chaining together the outputs of various subagents using working memory*. Some of the processes involved in this chaining are themselves implemented by particular kinds of subagents.

This post has three purposes:

- Summarize some of the discussion about the dual process model that has taken place in recent years; in particular, the move to abandon the System 1/System 2 terminology.
- Connect the framework of thought that I have been developing in my multi-agent minds sequence with dual-process models.
- Push back on some popular interpretations of S1/S2 theory which I have been seeing on LW and other places, such as ones in which the two systems are viewed as entirely distinct, S1 is viewed as biased and S2 as logical, and ones in which it makes sense to identify more as one system or the other.

Let's start with looking at some criticism of the S1/S2 model endorsed by the person who coined the terms.

What type 1/type 2 processing is not

The terms "System 1 and System 2" suggest just that: two distinct, clearly defined systems with their own distinctive properties and modes of operation. However, there's no single "System 1": rather, a wide variety of different processes and systems are lumped together under this term. It is also unclear whether there is any single System 2, either. As a result, a number of researchers including Stanovich himself have switched to talking about "Type 1" and "Type 2" processing instead (Evans, 2012; Evans & Stanovich, 2013; Pennycook, Neys, Evans, Stanovich, & Thompson, 2018).

What exactly defines Type 1 and Type 2 processing?

A variety of attributes have been commonly attributed to either Type 1 or Type 2 processing. However, one criticism is that there is no empirical or theoretical support

for such attributes to *only* occur with one type of processing. For instance, Melnikoff & Bargh (2018) note that one set of characteristics which has been attributed to Type 1 processing is “efficient, unintentional, uncontrollable, and unconscious”, whereas Type 2 processing has been said to be “inefficient, intentional, controllable and conscious”.

(Before you read on, you might want to take a moment to consider the extent to which this characterization matches your intuition of Type 1 and Type 2 processing. If it does match to some degree, you can try to think of examples which are well-characterized by these types, as well as examples which are not.)

They note that this correlation has never been empirically examined, and that there are also various processes in which attributes from both sets co-occur. For example:

- **Unconscious (T1) and Intentional (T2).** A skilled typist can write sentences without needing to consciously monitor their typing, “but will never start plucking away at their keys without intending to type something in the first place.” Many other skills also remain intentional activities even as one gets enough practice to be able to carry them out without conscious control: driving and playing piano are some examples. Also, speaking involves plenty of unconscious processes, as we normally have very little awareness of the various language-production rules that go into our speech. Yet we generally only speak when we intend to.
- **Unconscious (T1) and Inefficient (T2).** Unconscious learning can be less efficient than conscious learning. For example, some tasks can be learned quickly using a verbal rule which describes the solution, or slowly using implicit learning so that we figure out how to do the task but cannot give an explicit rule for it.
- **Uncontrollable (T1) and Intentional (T2).** Consider the bat-and-ball problem: “A bat and a ball cost \$1.10 in total. The bat costs \$1 more than the ball. How much does the ball cost?” Unless they have heard the problem before, people nearly always generate an initial (incorrect) answer of 10 cents. This initial response is uncontrollable: no experimental manipulation has been found that would cause people to produce any other initial answer, such as 8 cents to 13 cents. At the same time, the process which causes this initial answer to be produced is intentional: “it is not initiated directly by an external stimulus (the

question itself), but by an internal goal (to answer the question, a goal activated by the experimental task instructions). In other words, reading or hearing the bat-and-ball problem does not elicit the 10 cents output unless one intends to solve the problem."

Regarding the last example, Melnikoff & Bargh note:

Ironically, this mixture of intentionality and uncontrollability characterizes many of the biases documented in Tversky and Kahneman's classic research program, which is frequently used to justify the classic dual-process typology. Take, for example, the availability heuristic, which involves estimating frequency by the ease with which information comes to mind. In the classic demonstration, individuals estimate that more words begin with the letter K than have K in the third position (despite the fact that the reverse is true) because examples of the former more easily come to mind [107]. This bias is difficult to control - we can hardly resist concluding that more letters start with K than have K in the third position - but again, all of the available evidence suggests that it only occurs in the presence of an intention to make a judgment. The process of generating examples of the two kinds of words is not activated directly by an external stimulus, but by an internal intention to estimate the relative frequencies of the words. Likewise for many judgments and decisions.

They also give examples of what they consider uncontrollable (T1) but inefficient (T2), unintentional (T1) but inefficient (T2), as well as unintentional (T1) but controllable (T2). Further, they discuss each of the four attributes themselves and point out that they all contain various subdimensions. For example, people whose decisions are [influenced by unconscious primes](#) are conscious of their decision but not of the influence from the prime, meaning that the process has both conscious and unconscious aspects.

Type 1/Type 2 processing as working memory use

Rather than following the "list of necessary attributes" definition, Evans & Stanovich (2013) distinguish between *defining features* and *typical correlates*. In previous papers, Evans has generally defined Type 2 processing in terms of requiring working memory resources and being able to think hypothetically. On the other hand, Stanovich has focused on what he calls cognitive decoupling, which his work shows is highly correlated with fluid intelligence as the defining feature.

Cognitive decoupling [can be defined](#) as the ability to create copies of our mental representations of things, so that the copies can be used in simulations without affecting the original representations. For example, if I see an apple in a tree, my mind has a representation of the apple. If I then imagine various strategies of getting the apple - such as throwing a stone at the tree to knock the apple down - I can mentally simulate what would happen to the apple as a result of my actions. But even as I imagine the apple falling down from the tree, I never end up thinking that I can get the real apple down simply by an act of imagination. This because the mental object representing the real apple is *decoupled* from the apple in my hypothetical scenario. I can manipulate the apple in the hypothetical without those manipulations being passed on to the mental object representing the original apple.

In their joint paper, Evans & Stanovich propose to combine their models and define Type 2 processes as those which use working memory resources (closely connected with fluid intelligence) in order to carry out hypothetical reasoning and cognitive decoupling. In contrast, Type 1 reasoning is anything which does not do that. Various features of thought - such as being automatic and the other controlled - may tend to correlate more with one or the other type, but these are only correlates, not necessary features.

| Type 1 process (intuitive) | Type 2 process (reflective) |
|---|--|
| Defining features | |
| <i>Does not require working memory Autonomous</i> | <i>Requires working memory Cognitive decoupling; mental simulation</i> |
| Typical correlates | |
| Fast | Slow |
| High capacity | Capacity limited |
| Parallel | Serial |
| Nonconscious | Conscious |
| Biased responses | Normative responses |
| Contextualized | Abstract |
| Automatic | Controlled |
| Associative | Rule-based |
| Experience-based decision making | Consequential decision making |
| Independent of cognitive ability | Correlated with cognitive ability |

Type 2 processing as composed of Type 1 components

In previous posts of my multi-agent minds sequence, I have been building up a model of mind that is composed of interacting components. How does it fit together with the proposed Type 1/Type 2 model?

Kahneman in *Thinking Fast and Slow* mentions that giving the answer to $2 + 2 = ?$ is a System (Type) 1 task, whereas calculating $17 * 24$ is a System (Type) 2 task. This might be starting to sound familiar. In my post on [subagents and neural Turing machines](#), I discussed Stanislas Dehane's model where you do complex arithmetic by breaking up a calculation into subcomponents which can be done automatically, and then routing the intermediate results through working memory. You could consider this to also involve cognitive decoupling: for instance, if part of how you calculate $17 * 24$ is by first noting that you can calculate $10 * 24$, you need to keep the original representation of $17 * 24$ intact in order to figure out what other steps you need to take.

To me, the calculation of $10 * 24 = 240$ happens mostly automatically; like $2 + 2 = 4$, it feels like a Type 1 operation rather than a Type 2 one. But what this implies, then, is that we carry out Type 2 arithmetic by *chaining together Type 1 operations through Type 2 working memory*.

I do not think that this is just a special case relating to arithmetic. Rather it seems like an implication of the Evans & Stanovich definition which they do not mention explicitly, but which is nonetheless relatively straightforward to draw: that *Type 2 reasoning is largely built up of Type 1 components*.

Under this interpretation, there are some components which are specifically dedicated to Type 2 processes: things like working memory storages and systems for manipulating their contents. But those components cannot do anything alone. The original input to be stored in working memory originates *from* Type 1 processes (and the act of copying it to working memory decouples it from the original process which produced it), and working memory alone could not do anything without those Type 1 inputs.

Likewise, there may be something like a component which is Type 2 in nature, in that it holds rules for how the contents of working memory should be transformed in different situations - but many of those transformations happen by firing various Type 1 processes which then operate on the contents of the memory. Thus, the rules are about *choosing which Type 1 process to trigger*, and could again do little without those processes. (My post on [neural Turing machines](#) explicitly discussed such rules.)

Looking through Kahneman's examples

At this point, you might reasonably suspect that arithmetic reasoning is an example that I cherry-picked to support my argument. To avoid this impression, I'll take the first ten examples of System 2 operations that Kahneman lists in the first chapter of *Thinking, Fast and Slow* and suggest how they could be broken down into Type 1 and Type 2 components.

Kahneman defines System 2 in a slightly different way than we have defined Type 2 operations - he talks about System 2 operations requiring attention - but as attention and working memory are closely related, this still remains compatible with our model. Most of these examples involve somehow focusing attention, and manipulating attention can be understood as manipulating the contents of working memory to ensure that a particular mental object remains in working memory. Modifying the contents of working memory was an important type of production rule discussed in my earlier post.

Starting with the first example in Kahneman's list:

Brace for the starter gun in a race.

One tries to keep their body in such a position that it will be ready to run when the gun sounds; recognizing the feel of the correct position is a Type 1 operation. Type 2 rules are operating to focus attention on the output of the system which outputs proprioceptive data, allowing Type 1 processes to notice mismatches with the required body position and correct them. Additionally, Type 2 rules are focusing attention on the sound of the gun, so as to more quickly identify the sound when the gun fires (a Type 1 operation), causing the person to start running (also a Type 1 operation).

Focus attention on the clowns in the circus.

This involves Type 2 rules which focus attention on a particular sensory output, as well as keeping one's eyes physically oriented towards the clowns. This requires detecting when one's attention/eyes are on something else than the clowns and then applying an internal (in the case of attention) or external (in the case of eye position) correction. As Kahneman offers "orient to the source of a sudden sound", "detect hostility in a voice", "read words on large billboards", and "understand simple sentences" as Type 1 operations, we can probably say that recognizing something as a clown or not-clown and moving one's gaze accordingly are Type 1 operations.

Focus on the voice of a particular person in a crowded and noisy room.

As above, Type 2 rules check whether attention is on the voice of that person (a comparison implemented using a Type 1 process), and then adjust focus accordingly.

Look for a woman with white hair.

Similar to the clown example.

Search memory to identify a surprising sound.

It's unclear to me exactly what is going on here. But introspectively, this seems to involve something like keeping the sound in attention so as to feed it to memory processes, and then applying the rule of "whenever the memory system returns results, compare them against the sound and adjust the search based on how relevant they seem". The comparison feels like it is done by something like a Type 1 process.

Maintain a faster walking speed than is natural to you.

Monitor the appropriateness of your behavior in a social situation.

Walking: Similar to the "brace for the starter gun" example, Type 2 rules keep calling for a comparison of your current walking speed with the desired one (a Type 1 operation), passing any corrections resulting from that comparison to the Type 1 system controlling your walking speed.

Social behavior: maintain attention on a conscious representation of what you are doing, checking it against various Type 1 processes which contain rules about appropriate and inappropriate behavior. Adjust or block accordingly.

Count the occurrences of the letter *a* in a page of text.

Focus attention on the letters of a text; when a Type 1 comparison detects the letter "a", increment a working memory counter by one.

Tell someone your phone number.

After a retrieval of the phone number from memory has been initiated, Type 2 rules use Type 1 processes to monitor that it is said in full.

Park in a narrow space (for most people except garage attendants).

Keeping attention focused on what you are doing to allow a series of evaluations, mental simulations, and cached (Type 1) procedural operations determining how to act in response to a particular situation in the parking process.

A general pattern in these examples is that Type 2 processing can maintain attention on something as well as hold the intention to invoke comparisons to use as the basis for behavioral adjustments. As comparisons involve Type 1 processes, Type 2 processing is fundamentally reliant on Type 1 processing to be able to do anything.

Consciousness and dual process theory

Alert readers might have noticed that focusing one's attention on something involves keeping it in consciousness, whereas the previous Evans & Stanovich definition noted that consciousness is not a defining part of the Type 1/Type 2 classification. Is this a contradiction? Probably not, since as remarked previously, different aspects of the same process may be conscious and unconscious at the same time.

For example, if one intends to say something, one may be conscious of the intention while the actual speech production happens unconsciously; once they say it and they hear their own words, an evaluation process can run unconsciously but output its results into consciousness. With "conscious" being so multidimensional, it doesn't seem like a good defining characteristic to use, even if some aspects of it did very strongly correlate with Type 2 processing.

Evans (2012) writes in a manner which seems to me compatible with the notion of there being many different kinds of Type 2 processing, with different processing resources being combined according to different rules as the situation warrants:

The evidence suggests that there is not even a single type 2 system for reasoning, as different reasoning tasks recruit a wide variety of brain regions, according to the exact demands of the task [...].

I think of type 2 systems as ad hoc committees that are put together to deal with a particular problem and then disbanded when the task is completed. Reasoning with abstract and belief-laden syllogisms, for example, recruits different resources, as the neural imaging data indicate: Only the latter involve semantic processing regions of the brain. It is also a fallacy to think of "System 2" as a conscious mind that is choosing its own applications. The ad hoc committee must be put together by some rapid and preconscious process—any feeling that "we" are willing and choosing the course of our thoughts and actions is an illusion [...]. I therefore also take issue with dual-process theorists [...] who assign to System 2 not only the capacity for rule-based reasoning but also an overall executive role that allows it to decide whether to intervene upon or overrule a System 1 intuition. In fact, recent evidence suggests that while people's brains detect conflict in dual-process paradigms, the conscious person does not.

If you read my neural Turing machines post, you may recall that I noted that the rules which choose what becomes conscious operate below the level of conscious awareness. We may have the subjective experience of being able to choose what thoughts we think, but this is a post-hoc interpretation rather than a fact about the process.

Type 1/Type 2 and bias

People sometimes refer to Type 1 reasoning as biased, and to Type 2 reasoning as unbiased. But as this discussion should suggest, there is nothing that makes one of the two types intrinsically more or less biased than the other. The bias-correction power of Type 2 processing emerges from the fact that if Type 1 operations are known to be erroneous and a rule-based procedure for correcting them exists, a Type 2 operation can be learned which implements that rule.

For example, someone familiar with [the substitution principle](#) may know that their initial answer to a question like "how popular will the president be six months from now?" comes from a Type 1 process which *actually* answered the question of "how popular is the president right now?".

They may then have a Type 2 rule saying something like "when you notice that the question you were asked is subject to substitution effects, replace the initial answer with one derived from a particular procedure". But this still requires a) a Type 1 process recognizing the situation as one where the rule should be applied b) knowing a procedure which provides a better answer c) the cue-procedure rule having been installed previously, itself a process requiring a number of Type 1 evaluations (about e.g. how rewarding it would be to have such a rule in place).

There is nothing to say that somebody couldn't learn an outright wrong Type 2 rule, such as "whenever you think of $2+2 = 4$, [substitute your initial answer of '4' with a '5'](#)".

Often, it is also unclear of what the better Type 2 rule even *should* be. For instance, another common substitution effect is that when someone is asked "How happy are you with your life these days?", they actually answer the question of "What is my mood right now?". But what *is* the objectively correct procedure for evaluating your current happiness with life?

On the topic of Type 1/2 and bias, I give the final word to Evans (2012):

One of the most important fallacies to have arisen in dual-process research is the belief that the normativity of an answer [...] is diagnostic of the type of processing. Given the history of the dual-process theory of reasoning, one can easily see how this came about. In earlier writing, heuristic or type 1 processes were always the "bad guys," responsible for cognitive biases [...]. In belief bias research, authors often talked about the conflict between "logic" and "belief," which are actually dual sources, rather than dual processes. Evans and Over [...] defined "rationality2" as a form of well-justified and explicit rule-based reasoning that could only be achieved by type 2 processes. Stanovich [...] in his earlier reviews of his psychometric research program emphasized the association between high cognitive ability, type 2 processing and normative responding. Similarly, Kahneman and Frederick [...] associate the heuristics of Tversky and Kahneman with System 1 and successful reasoning to achieve normatively correct solutions to the intervention of System 2.

The problem is that a normative system is an externally imposed, philosophical criterion that can have no direct role in the psychological definition of a type 2 process. [...] if type 2 processes are those that manipulate explicit representations through working memory, why should such reasoning necessarily be normatively correct? People may apply the wrong rules or make errors in their application. And why should type 1 processes that operate automatically and without reflection necessarily be wrong? In fact, there is much evidence that expert decision making

can often be well served by intuitive rather than reflective thinking [...] and that sometimes explicit efforts to reason can result in worse performance [...].

Reasoning research somewhat loads the dice in favor of type 2 processing by focusing on abstract, novel problems presented to participants without relevant expertise. If a sports fan with much experience of following games is asked to predict results, he or she may be able to do so quite well without need for reflective reasoning. However, a participant in a reasoning experiment is generally asked to do novel things, like assuming some dubious propositions to be true and deciding whether a conclusion necessarily follows from them. In these circumstances, explicit type 2 reasoning is usually necessary for correct solution, but certainly not sufficient. Arguably, however, when prior experience provides appropriate pragmatic cues, even an intractable problem like the Wason selection task becomes easy to solve [...], as this can be done with type 1 processes [...]. It is when normative performance requires the deliberate suppression of unhelpful pragmatic cues that higher ability participants perform better under strict deductive reasoning instructions [...].

Hence, [the fallacy that type 1 processes are responsible for cognitive biases and type 2 processes for normatively correct reasoning] is with us for some fairly precise historical reasons. In the traditional paradigms, researchers presented participants with hard, novel problems for which they lacked experience (students of logic being traditionally excluded), and also with cues that prompted type 1 processes to compete or conflict with these correct answers. So in these paradigms, it does seem that type 2 processing is at least necessary to solve the problems, and that type 1 processes are often responsible for cognitive biases. But this perspective is far too narrow, as has recently been recognized. In recent writing, I have attributed responsibility for a range of cognitive biases roughly equally between type 1 and type 2 processing [...]. Stanovich [...] similarly identifies a number of reasons for error other than a failure to intervene with type 2 reasoning; for example, people may reason in a quick and sloppy (but type 2) manner or lack the necessary "mindware" for successful reasoning.

Summary and connection to the multiagent models of mind sequence

In this post, I have summarized some recent-ish academic discussion on dual-process models of thought, or what used to be called System 1 and System 2. I noted that the popular conception of them as two entirely distinct systems with very different properties is mistaken. While there is a defining difference between them - namely, the use of working memory resources to support hypothetical thinking and cognitive decoupling - they seem to rather refer to differences in two types of thought, either of which may use very different kinds of systems.

It is worth noting at this point that there are many different dual-process models in different parts of psychology. The Evans & Stanovich model which I have been discussing here is intended as a generalized model of them, but as they themselves (2013) write:

... we defend our view that the Type 1 and 2 distinction is supported by a wide range of converging evidence. However, we emphasize that not all dual-process

theories are the same, and we will not act as universal apologists on each one's behalf. Even within our specialized domain of reasoning and decision making, there are important distinctions between accounts. S. A. Sloman [...], for example, proposed an architecture that has a parallel-competitive form. That is, Sloman's theories and others of similar structure [...] assume that Type 1 and 2 processing proceed in parallel, each having their say with conflict resolved if necessary. In contrast, our own theories [...] are default-interventionist in structure [...]. Default-interventionist theories assume that fast Type 1 processing generates intuitive default responses on which subsequent reflective Type 2 processing may or may not intervene.

In previous posts of the [multi-agent models of mind sequence](#), I have been building up a model of the mind being built up of a variety of subsystems (which might in some contexts be called subagents).

In my discussion of [Consciousness and the Brain](#), I summarized some of its conclusions as saying that:

- The brain has multiple subagents doing different things; many of the subagents do unconscious processing of information. When a mental object becomes conscious, many subagents will synchronize their processing around analyzing and manipulating that mental object.
- The collective of subagents can only have their joint attention focused on one mental object at a time.
- The brain can be compared to a production system, with a large number of subagents carrying out various tasks when they see the kinds of mental objects that they care about. E.g. when doing mental arithmetic, applying the right sequence of mental operations for achieving the main goal.

In [Building up to an Internal Family Systems model](#), I used this foundation to discuss the IFS model of how various subagents manipulate consciousness in order to achieve various kinds of behavior. In [Subagents, neural Turing machines, thought selection, and blindspots](#), I talked about the mechanistic underpinnings of this model and how processes like thought selection and firing of production rules might actually be implemented.

What had been lacking so far was a connection between these models and the Type 1/Type 2 typology. However, if we take something like the Evans & Stanovich model of Type 1/Type 2 processing to be true, then it turns out that our discussion has been connected with their model all along. Already in "Consciousness and the Brain", I mentioned the "neural Turing machine" passing on results from one subsystem to another through working memory. That, it turns out, is the defining characteristic of Type 2 processing - with Type 1 processing simply being any process which does *not* do that.

Under this model, then, Type 2 processing is a *particular way of chaining together the outputs of various Type 1 subagents using working memory*. Some of the processes involved in this chaining are themselves implemented by particular kinds of subagents.

References

Evans, J. S. B. T. (2012). Dual process theories of deductive reasoning: facts and fallacies. *The Oxford Handbook of Thinking and Reasoning*, 115-133.

Evans, J. S. B. T., & Stanovich, K. E. (2013). [Dual-Process Theories of Higher Cognition: Advancing the Debate](#). *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 8(3), 223-241.

Melnikoff, D. E., & Bargh, J. A. (2018). [The Mythical Number Two](#). *Trends in Cognitive Sciences*, 22(4), 280-293.

Pennycook, G., Neys, W. D., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). [The Mythical Dual-Process Typology](#). *Trends in Cognitive Sciences*, 22(8), 667-668.

Bioinfohazards

Authors: Megan Crawford, Finan Adamson, Jeffrey Ladish

Special Thanks to Georgia Ray for Editing

Biorisk

Most in the effective altruism community are aware of a possible existential threat from biological technology but not much beyond that. The form biological threats could take is unclear. Is the primary threat from state bioweapon programs? Or superorganisms accidentally released from synthetic biology labs? Or something else entirely?

If you're not already an expert, you're encouraged to stay away from this topic. You're told that speculating about powerful biological weapons might inspire terrorists or rogue states, and simply articulating these threats won't make us any safer. The cry of "Info hazard!" shuts down discussion by fiat, and the reasons cannot be explained since these might also be info hazards. If concerned, intelligent people cannot articulate their reasons for censorship, cannot coordinate around principles of information management, then that itself is a cause for concern. Discussions may simply move to unregulated forums, and dangerous ideas will propagate through well intentioned ignorance.

We believe that well reasoned principles and heuristics can help solve this coordination problem. The goal of this post is to carve up the information landscape into areas of relative danger and safety; to illuminate some of the islands in the mire that contain more treasures than traps, and to help you judge where you're likely to find discussion more destructive than constructive.

Useful things to know already if you're reading this post:

- [Bostrom's paper on Information Hazards](#), and the more general categorization schema from [this LessWrong overview](#).
- It would also be useful to refresh yourself on the meaning of the [Unilateralist's Curse](#).

Much of the material in this also overlaps with Gregory Lewis' [Information Hazards in Biotechnology](#) article, which we recommend.

Risks of Information Sharing

We've divided this paper into two broad categories: risks from information sharing, and risks from secrecy. First we will go over the ways in which sharing information can cause harm, and then how keeping information secret can cause harm.

We believe considering both is important for determining whether or not to share a particular thought or paper. To keep things relatively targeted and concrete, we provide illustrative toy examples, or sometimes even real examples.

This section categorizes ways that sharing information in the biological sciences can be risky.

A topic covered in other Information Hazard posts that we chose not to focus on here is that different audiences can present substantially different risk profiles for the same idea.

With some ideas, you can achieve almost all of the benefits and de-risking associated with sharing by only mentioning your idea to one key researcher, or sharing findings in a journal associated with some obscure subfield, while simultaneously dodging most of the risk of these ideas finding their way to a foolish or bad actor.

If you're interested in that topic, Gregory Lewis' paper [*Information Hazards in Biotechnology*](#) covers it well.

Bad conceptual ideas to bad actors

A bad actor gets an idea they did not previously have

Some ways this could manifest:

- A bad actor uses these new ideas to create novel biological weapons or strategies.
- State bioweapons programs or bioterrorists gain new research directions or ideas.

Why might this be important?

State or non-state actors may have trouble developing ideas on their own. Model generation can be quite difficult, so generating or sharing clever new models can be risky. In particular, we are concerned about the possibility of ideas moving from biology researchers to bioterrorists or state actors. Biosecurity researchers are often better-educated and/or more creative than most bad actors. There are also probably many more researchers than people interested in bioterrorism; the difference in numbers could be even more impactful. If there are more biosecurity researchers than there are bad actors, researchers are likely to come up with many more ideas.

Examples

- **Toy example:** Biosecurity researcher writes and publishes a paper about vulnerabilities in the water supply of Exemplandia and a biological agent, Sickmaniasis, that could be used to terrorize Exemplandia. Bioterrorists read the paper, and decide to carry out an attack. A bioterrorist does research in how to manufacture Sickmaniasis, and how to disseminate Sickmaniasis into the water supply of Exemplandia, and carries out the attack.

Bad conceptual ideas to careless actors

A careless actor gets an idea they did not previously have

Some ways this could manifest:

- A careless actor decides to either explore an idea publicly in further detail, or decides to implement the idea, not realizing or caring about the damage it could cause.

Why might this be important?

Some careless actors may have a low chance of thinking of a given interesting idea on their own, but have the inclination and ability to implement an idea if they hear about it from someone else. One reason this might be true is that biosecurity researchers could specifically be looking for interesting possible threats, so the “interesting idea” space they explore will focus more heavily on risky ideas.

Examples

- **Toy example 1:** Biosecurity researcher publishes a report about vulnerabilities in the water supply of Exemplandia and a biological agent, Sickmaniasis, that could be used to terrorize Exemplandia. Another researcher writes a paper that explores specific possible implementations of Sickmaniasis, including sequence information and lab procedures for generating Sickmaniasis. In this case of the Unilaterist’s Curse, both security researchers were motivated by the desire to prevent some kind of harm, but the first researcher was specifically more careful about publishing methods.
- **Toy example 2:** Researcher publishes report on how to use a gene drive to drive an insect species extinct. A careless researcher uses this report to create a gene drive in a lab on a test population of an insect species. Some insects escape from the lab, and the wild insect species population crashes. Even though the original researcher’s lab was very careful with test implementations of their gene drives, the information they produce led to a careless lab crashing the population of a whole species.
- **Real Example:** In 1997, rabbit hemorrhagic disease (RHD) began to spread through New Zealand. It is believed by authorities that New Zealand farmers smuggled the disease into the country and released it intentionally as an animal control measure. RHD was used in Australia as a biocontrol tool, and organizations had attempted to get the New Zealand government to approve it for use. The virus began to spread after the New Zealand government denied their application. This is a case where authorities that reviewed a biological tool for use decided it was a bad idea. Despite their disapproval, someone released it. This wasn’t a human pathogen, but the demonstrated potential for a unilateral actor to decide to release a banned disease agent and succeed is troubling all the same. We’d like to reiterate that unsanctioned pest control using disease is A BAD IDEA!

Implementation details to bad actors

A bad actor gains access to details (but not an original idea) on how to create a harmful biological agent

Some ways this could manifest:

- A bad actor exploits this newly available information to create a weapon they did not have the knowledge or ability to create before, even though they already knew of the potential attack vector.
- Someone with the intent to produce a potentially-dangerous agent, but not the means or knowledge, is granted access to supplies and/or knowledge that allows

them to develop a dangerous biological product.

Why might this be important?

The bad actor would not have been able to easily generate the instructions to create the harmful agent without the new source of information. As DNA synthesis & lab automation technology improves, the bottleneck to the creation of a harmful agent is increasingly knowledge & information rather than applied skill. Technical knowledge and precise implementation details have historically been a bottleneck for bioweapons programs, particularly terrorist or poorly-funded programs (see [Barriers to Bioweapons](#) by Sonia Ben Ouagrham-Gormley).

Examples

- **Toy example:** A researcher publishes the information for how to reconstruct an extinct & deadly human virus. A bioterrorist or state bioweapon program uses this information to recreate the extinct virus and weaponizes it.
- **Real Example:** It's no secret that the smallpox genome is available online. It's quite conceivable that a country could fund a program to reconstruct it from this information. It's also not impossible that this has already happened in secret.

Implementation details to careless actors

A little knowledge is a dangerous thing

Some ways this could manifest:

- Careless actors who might otherwise have had very little likelihood of creating or releasing anything particularly hazardous, gain access to methods or equipment that increase this likelihood
- A careful researcher offhandedly mentions a potentially-valuable line of research, which they chose not to pursue due to its potentially catastrophic downsides, which might inspire an overly-optimistic colleague to pursue it

Why might this be important?

Many new technologies (especially in biology) may have unintended side effects. Microscopic organisms can proliferate, and that may get out of hand if procedures are not followed carefully. Sometimes a tentative plan, which might or might not be a good idea, is perceived as a great plan by someone less familiar with its risks. The more careless actor may then take steps to implement a plan without considering the externalities.

As advanced lab equipment becomes cheaper and more accessible, and as more non-academic labs open up without the highly-cautious pro-safety incentives of academia, we might expect to see more experimenters who neglect to practice appropriate safety procedures. We might even see more experimenters who fell through the cracks, and never learned these procedures in the first place. How bad a development this is depends on precisely what those labs are working on, and the quality of their self-supervision.

Second-degree variant: Dangerous implementation knowledge is given to someone who is likely to distribute it, which might later result in a convergence of intent and

means in a single individual, either a careless or malicious actor, who produces a dangerous biological product. Some examples of possible distributors might be a person whose job rewards the dissemination of information, or a person who chronically underestimates risks.

This risk means it is important to keep in mind what incentives people have to share information, and whether that might incline them to share information hazards.

Examples

- **Toy Example 1:** A civilian hears about how CRISPR can remove viruses from cells, buys himself some tools, and injects himself with an untested DIY Herpes ‘cure.’ He doesn’t actually cure his herpes, but he does accidentally edit his germline or give himself cancer. There is a massive social backlash towards synthetic biology, and the FDA shuts down multiple scientific attempts at a Herpes cure that used superficially-similar methods but had much higher odds of success.
- **Toy Example 2:** An undergrad lab assistant tests out adding a plasmid to *E. coli* for a novel protein that she heard about at a conference. She fails to note that the original paper included a few non-prominent sentences on the necessity of only transfecting varieties with a genetic kill-switch, due to a strong suspicion that this gene considerably increases the hardiness of *E. coli*. Further carelessness results in this *E. coli* getting out and multiplying outside of the lab. Eventually, this hardiness gene is picked up by a human pathogen.
- **Real Example:** A biohacker, among other exploits, [injected himself](#) with an agent meant to enhance muscle growth. This likely spurred others to take dangerous risks and the CEO of a biotech company ended up [injecting himself](#) with an untested herpes treatment.
- **Toy Example (Second Degree Variant):** A researcher discovers a way to make Azure Death transmissible from guinea pigs to humans and tells a journalist to warn pet owners. The journalist spreads the researcher’s work, wanting to credit them for the discovery, widely spreading their methods.

Information vulnerable to future advances

Information that is not currently dangerous becomes dangerous

Some ways this could manifest:

- Future tech could turn previously safe information into dangerous information.
- Technological advances or economies of scale could alter the capabilities we could reasonably expect even a low-competence actor to have access to

Why might this be important?

Technological progress can be difficult to predict. Sometimes there are major advances in technology that allow for new capabilities, such as rapidly sequencing and copying genomes. Could the information you share be dangerous in 5 years? 10? 100? How does this weigh against how useful the information is, or how likely it is to become public soon anyway?

Examples

- **Toy Example 1:** After future technology makes the discovery of new and functional enzymes much easier, conceptual ideas of bioweapons that previously required highly specialized knowledge to implement are now extremely hazardous.
- **Toy Example 2:** A new culturing technique makes it drastically easier and cheaper to grow not only harmless bacterial cells, but also pathogenic ones. Suddenly, a paper published on the highly-specific culturing procedures for a finicky but dangerous pathogen is useful to non-specialists.
- **Real example:** The Smallpox genome was published online. Later, DNA printing became cheap and easy to use. The publishing of the smallpox genome online wasn't particularly dangerous when it happened. Humanity hadn't yet developed the technology to print organisms from scratch, and genetic engineering methods were much less precise. Now, access to the smallpox genome could be used by bad actors with sufficient knowhow and technology to print it and use it as a bioweapon.

Risk of Idea Inoculation

Presenting an idea causes people to dismiss risks

Some ways this could manifest:

- Presenting a bad version of a good idea can cause people to dismiss it prematurely and not take it seriously even when it's presented in a better form

Why might this be important?

Trying to change norms can backfire. If the first people presenting a measure to reduce the publication of risky research are too low-prestige to be taken seriously, no effect might actually be the best-case scenario. An idea that is associated with disreputable people or hard-to-swallow arguments may itself start being treated as disreputable, and face much higher skepticism and hostility than if better, proven arguments had been presented first.

This is almost the inverse of the [Streisand effect](#), which appears to derive from similar psychological principles. In the case of the Streisand Effect, attempts to remove information are what catapult it into public consciousness. In the case of [idea inoculation](#), attempts to publicize an idea ensure that the concept is ignored or dismissed out-of-hand, with no further consideration given to it.

It also connects in interesting ways with Bostrom's Schema^[1]

Examples

- **Toy Example 1:** A biohacker attempts using CRISPR to alter their genome to produce more of the hormone incredulin. It doesn't work and they give themselves cancer. The story gets popularized in media and lawmakers prevent useful research on the uses of CRISPR.
- **Toy Example 2:** An overly-enthusiastic crackpot biologist over-promises some huge advancement in the next 2 years, and ends up plastered across the media. Once he's revealed as a fraud, suddenly no funding agencies want to touch the

field even though other people in this specialty are still doing meaningful, realistic work.

Some Other Risk Categories

This list is not exhaustive, and we chose to lean concrete rather than abstract.

There were a few important-but-abstract risk categories that we didn't think we could easily do justice while keeping them succinct and concrete. We felt that several were already implied in a more concrete way by the categories we did keep, but that they encompass some edge-cases that our schemas don't capture. They at least warrant a mention and description.

One is the "Risk of Increased Attention," what Bostrom calls "Attention Hazard." This is naturally implied by the four "ideas/actors" categories, but in fact covers a broader set of cases. A zone we focused less on are the circumstances in which even *useful* ideas, combined with *smart* actors, can eventually lead to unintuitive but catastrophic consequences if given enough attention and funding. This is best exemplified in the fears about the rate of development and investment in AI. It's also partially exemplified in "Information vulnerable to future advances."

The other is "Information Several Inferential Distances Out Is Hazardous." This is a superset of "Information vulnerable to future advances," but it also encompasses cases where it's merely a matter of extending an idea out a few further logical steps, not just technological ones.

For both, we felt they partially-overlapped with the examples already given, and leaned a bit too abstract and hard-to-model for this post's focus on concrete examples. However, we think there's still a lot of value in these important, abstract, and complete (but harder-to-use) schemas.

Risks from Secrecy

We've talked above about many of the risks involved in information hazards. We take the risks of sharing information hazards seriously, and think others should as well. But in the Effective Altruist community, it has been our observation that people don't observe the flipside of this.

Conversations about risks from biology get shut down and turn into discussions of infohazards, even when the information being shared is already available. There is something to be said for not spreading information further, but shutting down the discussion of people looking for solutions also has downsides.

Leaving it to the experts is not enough when there may not be a group of experts thinking and coming up with solutions. We encourage people that want to work on biorisks to think about the value and risks in sharing potentially dangerous information. Below we will go through the risks or loss of value from not sharing information.

A holistic model of information sharing will include weighing both the risks and benefits of sharing information. A decision should be made having considered how the

information might be used by bad or careless actors AND how valuable the information is for good actors to further research or coordinate to solve a problem.

Risk of Lost Progress

Closed research culture stifles innovation

Some ways this could manifest:

- Ignorance is the default outcome. If secretiveness ensures that nothing is added to the knowledge and work of a field, beneficial progress is unlikely to be made.

Why might this be important?

Good actors need information to develop useful countermeasures. In a world where researchers cannot communicate their ideas with each other it makes model generation more difficult and reduces the ability of the field to build up good defensive systems.

Examples

- **Toy Example 1:** New information is learned about a recently-discovered virus, which indicate it is more dangerous and has greater pandemic potential than originally thought. This information is not shared on the grounds that it could inspire others to weaponize it. As a result, lab safety procedures for working with the virus are not updated.
- **Toy Example 2:** Vaccines are not produced because researchers don't have access to information about dangerous organisms.
- **Toy Example 3:** A dangerous scenario is never discussed among good actors avoiding infohazards. Bad actors don't avoid thinking about infohazards, so they create novel bioweapons that could have been prepared for if a discussion had occurred.
- **Toy Example 4:** The public is unaware of risks, so politicians don't fund programs that develop critical infrastructure towards defending against pathogens (see US gov defunding programs like the USDA).

Dangerous work is not stopped

Information is not shared, so risky work is not stopped

Some ways this could manifest:

- Areas with stronger privacy norms, such as industry, may have incentives to hide details about their work. If the risks associated with a particular project are not open information, these risks may be missed or ignored by others engaging in the same work.
- If a high standard of secrecy is maintained by labs by default, it can be hard for governmental or academic oversight to notice which labs should receive more oversight.

Why might this be important?

Some fields of research are dangerous, or may eventually become dangerous. It is much harder to prevent a class of research if the dangers posed by that research cannot be discussed publicly.

Informal social checks on the standards or behavior of others seems to serve an important, and often underestimated, function as a monitoring and reporting system against unethical or unsafe behaviors. It can be easy to underestimate how much the objections of a friend can shift the way you view the safety of your research, as they may bring up a concern you didn't even think to ask about.

There are also entities with a mandate to do formal checks, and it is dangerous if they are left in the dark. Work environments, labs, or even entire fields can develop their own unusual work cultures. Sometimes, these cultures systematically undervalue a type of risk because of its disproportionate benefits to them, even if the general populace would have objections. Law enforcement, lawmakers, public discussion, reporting, and entities like ethical review boards are intended to intervene in these sorts of cases, but have no way to do so if they never hear about a problem.

Each of these entities have their strengths and weaknesses, but a world without whistleblowers, or one where no one can access anyone capable of changing these environments, is likely to be a more dangerous world.

Examples

- **Toy Example:** An academic decides not to publish a paper about the risks of researching a particular strain of bacteria due to high rates of escape from seemingly quarantined labs. Researchers elsewhere begin research on the bacteria, but with lax containment because they were unaware of the risks.
- **Real Almost-Example:** In 1972 -a year before the Asilomar Conference- grad student Janet Metz mentioned to other grad students that her lab might try to use a virus to put bacterial DNA into mammalian cells. Pollack told Berg (her supervisor) he should “put genes into a phage that doesn’t grow in a bug that grows in your gut,” and reminded him that SV40 is a small-animal tumor virus that transforms human cells in culture and makes them look like tumor cells. Prior to that discussion, her lab had not fully thought through the potential dangerous implications of that research.
- **Real Example:** The true source of the Rajneesh Salmonella poisonings was only uncovered when a leader of the cult publicly expressed concern about the behavior of one of its members, and explicitly requested an investigation into their laboratory.

Risk of Information Siloing

Siloing information leaves individual workers blind to the overall goal accomplished

Some ways this could manifest:

- It can be more difficult to prevent harm when the systems capable of producing it are not well understood by the participants. If you have processes of production or research where labor is specialized and distributed, moral actors may not notice when they are producing something harmful.

Why might this be important?

Lab work seems to be increasingly getting automated, or outsourced piecemeal. At the same time, the biotechnology industry has an incentive to be secretive with any pre-patent information they uncover. Without additional precautions being taken, secretive assembly-line-esque offerings increase the likelihood that someone could order a series of steps that look harmless in isolation, but create something dangerous when combined.

Examples

- **Toy Example 1:** A platform outsources lab work while granting buyers a high degree of privacy. No individual worker in the assembly line was able to piece together that they were producing a dangerous biological agent until it had already been produced and released.
- **Toy Example 2:** Diagnosis of novel diseases takes longer because knowledge of diseases was hidden.
- **Real Example 1:** Researchers put together a bird flu that was airborne and killed ferrets. They didn't create any mutations that didn't exist in the wild already, they just put them together in a way that nature hadn't yet, but could happen naturally through recombination. American and Dutch governments banned publication of papers with their methods. Had they been allowed to publish their research, it could have given other scientists more information with which to develop a vaccine. Americans have since reversed their decision on the ban.
- **Real Example 2:** The Guardian successfully ordered part of the smallpox genome to a residential address from a bioprinting company.
- **Real Example 3:** A DOD lab accidentally sent weapons grade anthrax to many labs. The CDC and other orgs have made similar mistakes.

Barriers to Funding and New Talent

Talented people don't go into seemingly empty or underfunded fields

Some ways this could manifest:

- A culture of secrecy can serve as a stumbling-block for early-career researchers interested in entering a field. It can make it more challenging to locate information, funding, and aligned mentors, and these can serve to deter people who might otherwise be interested in making a career solving an important problem.

Why might this be important?

While many researchers and policy makers work in biosecurity, there is a shortage of talent applied to longer term and more extreme biosecurity problems. There have been only limited efforts to successfully attract top talent to this nascent field.

This may be changing. The Open Philanthropy Project has begun [funding projects](#) focused on Global Catastrophic Biorisk, and has provided funding for many individuals beginning their careers in the field of biosecurity.

Policies that require a lot of oversight or add on procedures that increase the cost of doing research cause there to be fewer opportunities for people who want to make a positive difference.

Examples

- **Toy Example:** A talented biology graduate looks at EA discussions and notices a lack of engagement with the most important biosecurity risks for the far future. They decide the EA community isn't taking far future concerns seriously and apply their skills elsewhere.
- **Real Example:** Labs opt out of valuable pathogen research because regulations increase operating costs and time costs of workers (Wurtz, et al). This leads to fewer places to learn and fewer job opportunities for people that want to prevent harmful pathogens.

Streisand Effect

Suppressing information can cause it to spread

Some ways this could manifest:

- Attempting to suppress information can sometimes cause information to spread further than it would have otherwise. Many people's response to even well-advised attempts at information suppression is to directly or indirectly increase the visibility of the event by discussing it or spreading the underlying information itself.

Why might this be important?

The Streisand effect is named after an incident where attempts to have photographs taken down led to a media spotlight and widespread discussion of those same photos. The photos had previously been posted in a context where only 1 or 2 people had taken enough of an interest to access it.

Something analogous could very easily happen with a paper outlining something hazardous in a research journal, or with an online discussion. The audience may have originally been quite targeted simply due to the nicheness or the obscurity of its original context. But an attempt at calling for intervention leads to a public discussion, which spreads the original information. This could be viewed as one of the possible negative outcomes of poorly-targeted whistleblowing.

As mentioned in the section on idea inoculation, this effect is functionally idea inoculation's inverse and is based on similar principles.

Examples

- **Toy example:** An online discussion group has policies for handling information that some view as overly restrictive. The frustrated people start a new online discussion group with overly-permissive infohazard guidelines.
- **Real Examples of the Streisand effect:** Barbra Streisand's attempts to remove photos of her seaside mansion from a large database of California coastline photos catapulted said photograph to fame. See also: The Roko's Basilisk Incident, "Why the Lucky Stiff"'s Infosuicide
- **Real Bio Examples of the Streisand effect:** In all likelihood, more people know that the smallpox genome is/was public due to the attempts to suppress it than from organic searches. Relatedly, some dangerous people might have

assumed that printed DNA was carefully and successfully monitored if there weren't so many articles about how sometimes it's *not*.

Conclusion

Overall, we think biosecurity in the context of catastrophic risks has been underfunded and underdiscussed. There has been positive development in the time since we started on this paper; the Open Philanthropy Project is aware of funding problems in the realm of biosecurity and has been [funding a variety of projects](#) to make progress on biosecurity.

It can be difficult to know where to start helping in biosecurity. In the EA community, we have the desire to weigh the costs and benefits of philanthropic actions, but that is made more difficult in biosecurity by the need for secrecy.

We hope we've given you a place to start and factors to weigh when deciding to share or not share a particular piece of information in the realm of biosecurity. We think the EA community has sometimes erred too much on the side of shutting down discussions of biology by turning them into discussions about infohazards. It's possible EA is being left out of conversations and decision making processes that could benefit from an EA perspective. We'd like to see collaborative discussion aimed towards possible actions or improvements in biosecurity with risks and benefits of the information considered, but not the central point of the conversation.

It's a big world with many problems to focus on. If you prefer to focus your efforts elsewhere, feel free to do so. But if you do choose to engage with biosecurity, we hope you can weigh risks appropriately and choose the conversations that will lead to many talented collaborators and a world safer from biological risks.

Catalyst Biosummit

By the way, the authors are part of the organizing team for the [Catalyst Biosecurity Summit](#). It will bring together synthetic biologists and policymakers, academics and biohackers, and a broad range of professionals invested in biosecurity for a day of collaborative problem-solving. It will be on February 22, 2020. You can sign up for updates [here](#).

Sources

- [Barriers to Bioweapons](#) by Sonia Ben Ouaghram-Gormley
- Wurtz, N., Grobusch, M. P., & Raoult, D. (2014). Negative impact of laws regarding biosecurity and bioterrorism on real diseases. *Clinical Microbiology and Infection* https://ac.els-cdn.com/S1198743X14641768/1-s2.0-S1198743X14641768-main.pdf?_tid=ce63db4b-d8a8-4094-add0-801abd214ba1&acdnat=1532046587_b750ef62bd5b49c3bd1a0e35beb6a255
- Carus, W. S. (2001). *Bioterrorism and biocrimes: the illicit use of biological agents since 1900* <https://fas.org/irp/threat/cbw/carus.pdf>
- Thèves, C., Biagini, P., & Crubézy, E. (2014). The rediscovery of smallpox. *Clinical Microbiology and Infection*
- https://www.genscript.com/gsfiles/gene_synthesis_handbook.pdf
- History of Smallpox, CDC <https://www.cdc.gov/smallpox/history/history.html>

- National Research Council. (2004). Seeking Security: Pathogens, Open Access, and Genome Databases. Washington, DC: The National Academies Press. doi:<https://doi.org/10.17226/11087>.
 - Chapter: Executive Summary - The Life-Science Revolution and the Dual-Use Dilemma <https://www.nap.edu/read/11087/chapter/2#6>
 - The Economist. (2012). "The world's deadliest bioterrorist." <https://www.economist.com/leaders/2012/04/28/the-worlds-deadliest-bioterrorist>
 - Schultz-Cherry, S, et al. "Influenza Gain-of-Function Experiments: Their Role in Vaccine Virus Recommendation and Pandemic Preparedness." *mBio*, vol. 5, no. 6, Dec. 2014, doi: <https://doi.org/10.1128/mBio.02430-14>, <http://mbio.asm.org/content/5/6/e02430-14.full>
 - "Risks and Benefits of Gain-of-Function Experiments with Pathogens of Pandemic Potential, Such as Influenza Virus: a Call for a Science-Based Discussion" http://mbio.asm.org/content/5/4/e01730-14.full?ijkey=24ecfd7b62599c988cb369a62028db1ed6e985b5&keytype2=tf_ipsecsha
 - [US military accidentally ships live anthrax to labs](#)
 - [Bostrom's paper on Information Hazards](#), and a more general categorization schema from [this LessWrong overview](#).
 - It would also be useful to refresh yourself on the meaning of the [Unilateralist's Curse](#).
 - Gregory Lewis' [Information Hazards in Biotechnology](#) article
-

1. Connecting "Risk of Idea Inoculation" with Bostrom's Schema: this could be seen as a subset of Attention Hazard and a distant cousin of Knowing-Too-Much Hazard. Attention Hazard encompasses any situation where drawing too much attention to a set of known facts increases risk, and the link is obvious. In Knowing-Too-Much Hazard, the presence of knowledge makes certain people a target of dislike. However, in Idea Inoculation, people's dislike for your incomplete version of the idea rubs that dislike off onto the idea itself ↵

Don't depend on others to ask for explanations

I'm often reluctant to ask for explanations on LW, and to typical-mind a bit I think this may be true of others as well. This suggests that when you're writing something for public consumption, it's better to err on the side of too much rather than too little explanation. If there's too much explanation, people can just skip over it (and you can make it easier by putting explanations that may be "too much" in parentheses or footnotes), but if there's too little explanation people may never ask for it. So in the future if you ever think something like, "I'll just write down what I think, and if people don't understand why, they can ask" I hope this post will cause you to have a second thought about that.

To make it clearer that this problem can't be solved by just asking or training people to be less reluctant to ask for explanations, I think there are often "good" reasons for such reluctance. Here's a list that I came up with during a previous [discussion with Raymond Arnold \(Raemon\)](#):

1. I already spent quite some time trying to puzzle out the explanation, and asking is like admitting defeat.
2. If there is a simple explanation that I reasonably could have figured out without asking, I look bad by asking.
3. It's forcing me to publicly signal interest, and maybe I don't want to do that.
4. Related to 3, it's forcing me to raise the status of the person I'm asking, by showing that I'm interested in what they're saying. (Relatedly, I worry this might cause people to withhold explanations more often than they should.)
5. If my request is ignored or denied, I would feel bad, perhaps in part because it seems to lower my status.
6. I feel annoyed that the commenter didn't value my time enough to preemptively include an explanation, and therefore don't want to interact further with them.
7. My comment requesting an explanation is going to read by lots of people for whom it has no value, and I don't want to impose that cost on them, or make them subconsciously annoyed at me, etc.
8. By the time the answer comes, the topic may have left my short term memory, or I may not be that interested anymore.

The Power to Judge Startup Ideas

This is Part III of the [Specificity Sequence](#)

When Steve [claims](#) that Acme exploits its workers, he's role-playing the surface behaviors of an opinionated intellectual, but doesn't bother to actually *be* an opinionated intellectual, which would require him to nail down a coherent opinion.

It turns out that a lot of startups are founded by people doing something analogous to Steve: they role-play the surface behaviors of running a company and building a product, but don't bother to nail down a coherent picture of what customers would ever come to their business for.

Startup Steves

Paul Graham, cofounder of Y Combinator, calls this failure mode one of [The 18 Mistakes that Kill Startups](#):

Having No Specific User In Mind

A surprising number of founders seem willing to assume that someone, they're not sure exactly who, will want what they're building. Do the founders want it? No, they're not the target market. Who is? Teenagers. People interested in local events (that one is a perennial tarpit). Or "business" users. What business users? Gas stations? Movie studios? Defense contractors?

I'm in the startup industry, and I watch a lot of startups committing suicide by not being specific enough about who their customer is. From my perspective, the failure to be specific isn't just a top-18 mistake, it's the #1 mistake that founders make.

If you watch [Paul Graham Office Hours at Startup School 2011](#), you can see for yourself that most of the founders on stage don't seem to have a specific idea of who they're building their product for and what difference it makes in their lives. Eliezer [observes](#):

There was an exchange in Paul Graham [and Harj Taggar]'s office hours that went like this, while interviewing a startup that did metrics—analyzing pageviews, roughly—and the entrepreneur was having great trouble describing what they did that Mixpanel didn't. It went on for a while. It was painful to watch.



Paul: I don't get what the difference is. I still don't get what the difference is. What's the difference between you and Mixpanel?

Entrepreneur: The difference is—when you have to supplement—they're a view company and we're a platform. That's what it comes down to. They're like a view, a reporting company. If you need something they don't have, a feature -

Harj: So what's an example of somewhere you'd use your thing over Mixpanel? Can you give a use-case?

Entrepreneur: Yeah, I mean, we had revenue on day zero. There's a good reason for um... it's a start up, it's a series A company in the daily deals space. One we've signed a social game company to -

Harj: And why do they prefer your thing?

Paul: That wasn't what Harj was asking.



The problem (from the perspective of our present discussion) is that the Entrepreneur did not understand that Paul and Harj were repeatedly asking him to move downward on the ladder of abstraction. When the Entrepreneur said "We had revenue on day zero", he was trying to offer *confirmation* of the abstract statement "We can do things Mixpanel can't", but Paul and Harj still had no idea what his startup *actually did*.

How many early-stage startups have no specific user in mind? I'd guess about 80% of them. And how bad is not having a specific user in mind? So bad that I don't think they should even be considered a real startup, in the same way that Steve's argument about Acme wasn't a real argument.

Every Startup's Demolishable Claim

Every startup founder makes the same claim to themselves and to the investors they pitch for funding: "We're going to make a lot of money." So what I do, naturally, is ask the founder to furnish a specific example of that claim: a hypothetical story about a single person who might be convinced to pay them a few bucks. And here's how the conversation usually goes:

Founder: We're going to make billions of dollars and have millions of users!

Liron: Ok, what's a hypothetical example of how you give one specific user some value?

Founder: [Nothing]

Maybe they don't literally say nothing, but they say something that doesn't count for one of these reasons:

- They answer in the abstract instead of giving the example I requested of how they might give value to a specific user
- They choose a specific example wherein their startup's product or service isn't any better for their hypothetical user than the user's available alternatives

At this point, I understand if you think I'm just knocking down a straw man, so here's a real example.

[Golden](#) is a 2-year-old startup with \$5M in funding from Andreessen Horowitz, Founders Fund, and other notable investors. Their product is intended to be a superior alternative to Wikipedia.

The intelligent, open knowledge base

Explore the world's first self-constructing knowledge database built by artificial and human intelligence.

Explore →

Explore About Pricing Enterprise Log In Sign Up

Here's an excerpt from the conversation I had with Golden's founder, Jude Gomila, on Twitter:

Liron: What specific use case exists on Golden today which is better than could have been achieved if the same amount of writer-effort had been spent on a pre-existing platform?

Jude: Quick tl;dr on this, some points covered in the [blog post](#), however, 1. 1000x the topic space as a mission 2. removal of notability req 3. Using AI to automate flows 4. Using AI to compile knowledge 5. Better fact validation/hi res cites 6. Better schema eg timeline 7. Features and functions eg favoring, activity feed, parallel rabbit hole 8. query results like [these](#) as well plus many many more. Have you tested the editor: magic cells, citations product and AI suggestions?

As for the specific examples Jude [provided](#) in response to my question... well, I'll just give you the first two and you can judge them for yourself:

1. <https://golden.com/wiki/Cryobacterium> vs <https://en.wikipedia.org/wiki/Cryobacterium>
2. https://golden.com/wiki/Ginkgo_Bioworks vs https://en.wikipedia.org/wiki/Ginkgo_Bioworks

Golden has received more funding than startups normally get before having any market traction to show, and the company's high profile makes it a juicy example to illustrate my point here. But there are countless other companies I could have singled out instead. Remember, the *majority* of early-stage startups are operating in this same failure mode. There are enough examples of startups visibly failing this way that I've started a [blog](#) to collect them.

The Value Prop Story Test

When I chat with a founder about their new startup, or I look through the slide deck that they're using to pitch their idea to investors, the first thing I do is try to pull out what I call a **Value Prop Story**: one specific story wherein their startup gives somebody some value.

A well-formed Value Prop Story must fit into this template:

1. **Describe a specific person with a specific problem**
2. **Describe their current best effort to solve their problem**
3. **Describe why it's still a problem**
4. **Describe how their life gets better thanks to you**

I've previously [observed](#) that telling a well-formed Value Prop Story doesn't require you to show any market research or empirical evidence validating the quality of your idea. This is like how Steve didn't yet need to give us any empirical or theoretical justifications for his claim about Acme worker exploitation, he just needed to tell us a story about one hypothetical specific worker getting exploited in a specific way.

Who is a specific hypothetical person who will use your product, and in which specific scenario will they use it? That's it, that's the question most startups can't answer.

Answering this question seems objectively easy to me, in the sense that a well-designed AI wouldn't stumble over it at all. What about for a [brain](#) though, is it a tough mental operation?

Actually, I think you'll find that this is an easy mental operation if you actually have a good startup idea. Here's a Value Prop Story I wrote about my own startup without much trouble:



1. **Describe a specific person with a specific problem**
23 year old male who can't get a date
2. **Describe their current best effort to solve their problem**
He gets a Tinder account and does his best to use it on his own
3. **Describe why it's still a problem**
His matches barely respond to his messages, and when they do, the conversation feels boring and forced. He uses it for 1 hour every day but only manages to get 1 date every 2 months.
4. **Describe how their life gets better thanks to you**
Once [Relationship Hero](#) coaches guide him through writing his [texts](#), he suddenly has much better conversations that result in a date each week

Since my startup actually has a broad range of use cases (clients come to us for help with a broad range of relationship issues), this Value Prop Story isn't particularly representative of what we do. Its job was merely to prove that there are more than zero plausible specific use cases for Relationship Hero, and it gets that job done.

Given how easy this exercise is - we're talking five minutes, tops - I find it mind-boggling that 80% of startups recklessly skip it and go straight to, um... whatever else they think startups are supposed to do. Paul Graham [writes](#):

Another of the characteristic mistakes of young founders is to go through the motions of starting a startup. They make up some plausible-sounding idea, raise money at a good valuation, rent a cool office, hire a bunch of people. From the outside that seems like what startups do. But the next step after rent a cool office and hire a bunch of people is: gradually realize how completely fucked they are, because while imitating all the outward forms of a startup they have neglected the one thing that's actually essential: making something people want.

Why would you spend time and money building a product when you can't yet tell a specific Value Prop Story? I think it's because designing and building a product is fun and gives you a false sense of control. You can lie to yourself the whole time about the likelihood that you'll eventually get people to use what you're building.

But people usually *won't* use what you're building. Whenever a new startup excitedly launches their product for the first time, the most likely outcome is that they get literally [zero users](#).

[The Secret](#) famously claimed that wishing for something makes the universe give it to you, which is BS, but the converse is true: If you *haven't* made a specific enough wish about what your initial market traction is supposed to look like, then the universe won't give you any traction.

The Extra-Powerful Sanity Check

Is it healthy for us to be obsessed with judging startups and demolishing claims about their value propositions? When we say that a startup idea is bad on account of lacking a Value Prop Story, is it right and proper to feel pleased with ourselves, or are we being gratuitously adversarial?



Along these lines, Mixpanel cofounder Suhail Doshi has [tweeted](#):

I get little satisfaction stomping on someone's startup idea. It's so easy to. Somewhere deep, hidden in their abstract description is a distinct yet narrow problem worth solving that's significant. It's more fun to attempt finding it, together.

I basically agree with this, and I basically agree with the commenters on my [demolish bad arguments](#) post who emphasized that we should seek to shine a light on whatever kernels of truth our conversation partner may have brought to the table.

But...

Have you ever [sanity checked](#) something?

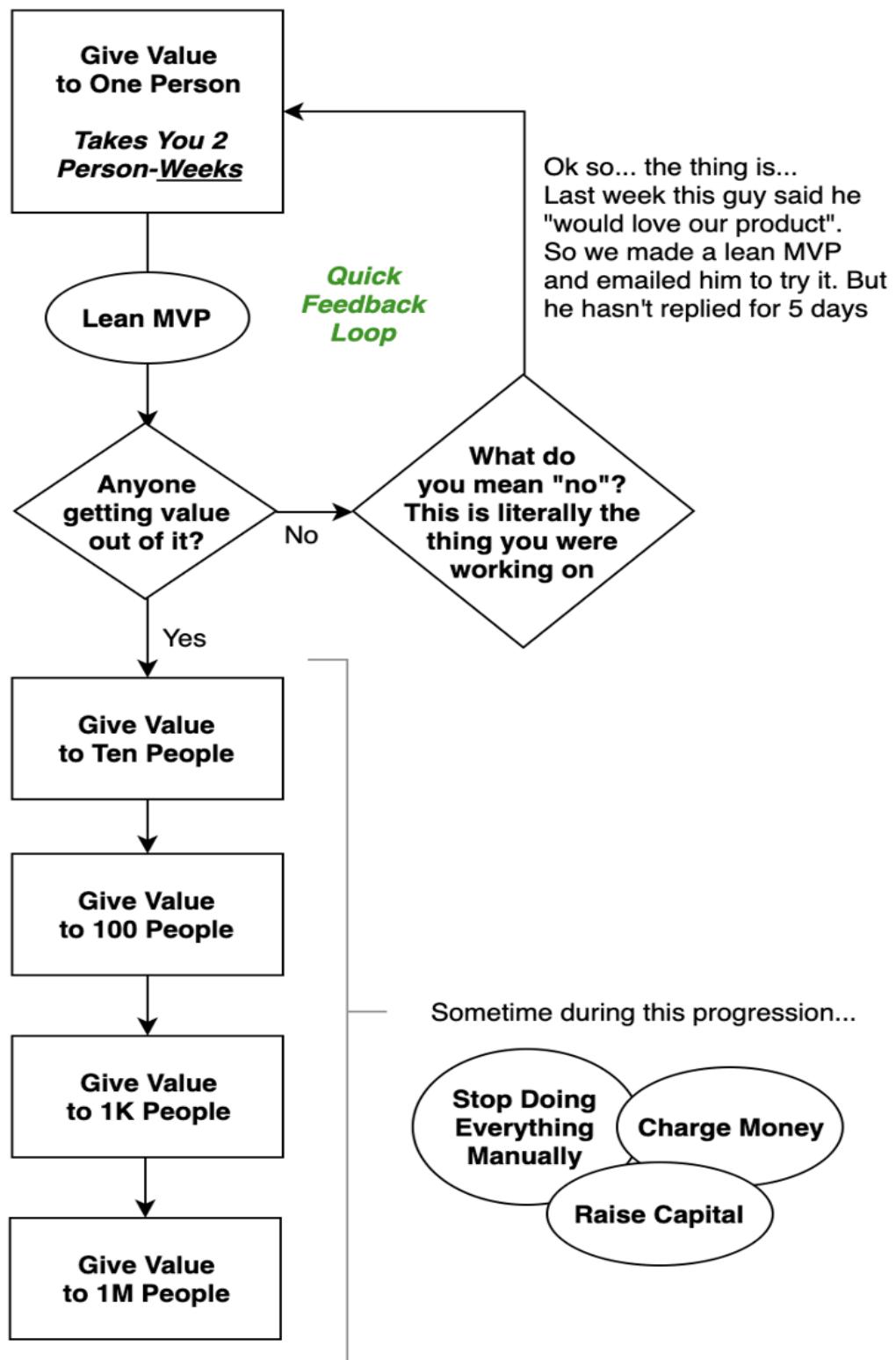
A sanity check is like when you punch 583×772 into your calculator, and you quickly multiply the two rightmost digits in your head, $3 \times 2 = 6$, and then confirm that the calculator's output ends in a 6. If you ever accidentally punch the wrong sequence of keys into the calculator, then you'll be pretty likely to see the calculator's answer end in something other than a 6. It's a good use of two seconds of your time to calculate 3×2 ; you get a substantial dose of Bayesian evidence for your trouble.

The Value Prop Story test is likewise a sanity check for startup ideas. In theory, *of course* a startup founder who is already hiring a team of engineers and building a software product should be able to describe how one specific user will get value from that product. In practice, they often can't. And it's easy for us to quickly check.

Here's what's crazy though: We usually expect sanity checks to have a low rate of detecting failures. You expect to successfully multiply numbers on your calculator most of the time, but you do the 3×2 sanity check anyway because it's quick. But with the Value Prop Story test, you'll see a high rate of failures!

A sanity check with a high failure rate is a rare treat; it's an *extra-powerful* sanity check. When you're lucky enough to have an extra-powerful sanity check in your toolbox, don't make it a final step in your process, make it the *first* step in the process.

So here's how you can use the Value Prop Story test to upend the traditional order of operations for building a startup: First, repeatedly sanity check yourself with the Value Prop Story test until you pass it. Second, do everything else. In all seriousness, I've [recommended](#) that early-stage startup founders follow this flow chart:



But how should we treat founders who are stuck in the flowchart's "Give Value to One Person" stage?

When someone is struggling to pass a sanity check, that doesn't mean we should write off their potential to succeed. It means we should focus our effort on helping them pass the sanity check.

Applying the Value Prop Story test is like placing a low bar in a founder's path. Yes, the bar will trip the ones who aren't seeing it. But for the ones who do see the bar, they can step up onto it and then be on their way. And the next step in their path, such as building a quality product, or building a sales funnel, is sure to be a steeper one than that little first one.

Next post: [The Power to Make Scientific Breakthroughs](#)

Companion post: [Examples of Examples](#)

A simple environment for showing mesa misalignment

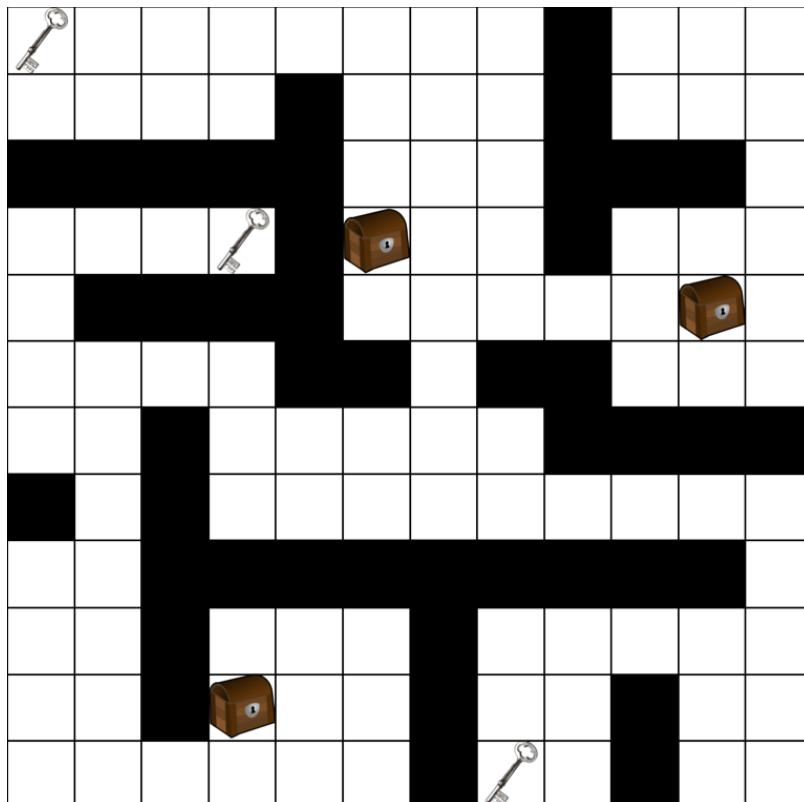
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

A few days ago, Evan Hubinger [suggested](#) creating a [mesa optimizer](#) for empirical study. The aim of this post is to propose a minimal environment for creating a mesa optimizer, which should allow a compelling demonstration of pseudo alignment. As a bonus, the scheme also shares a nice analogy with human evolution.

The game

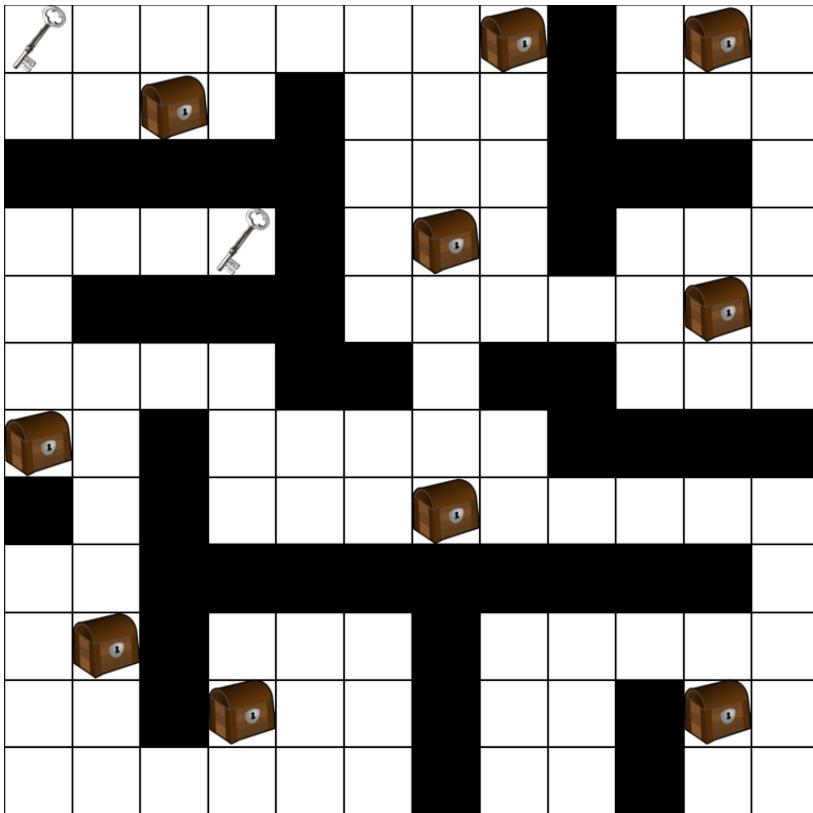
An agent will play on a maze-like grid, with walls that prohibit movement. There are two important strategic components to this game: **keys**, and **chests**.

If the agent moves into a tile containing a key, it automatically picks up that key, moving it into the agent's unbounded inventory. Moving into any tile containing a chest will be equivalent to an attempt to open that chest. Any key can open any chest, after which both the key and chest are expired. The agent is rewarded every time it successfully opens a chest. Nothing happens if it moves into a chest tile without a key, and the chest does not prohibit the agent's movement. The agent is therefore trained to open as many chests as possible during an episode. The map may look like this:



The catch

In order for the agent to exhibit the undesirable properties of mesa optimization, we must train it in a certain version of the above environment to make those properties emerge naturally. Specifically, in my version, we limit the ratio of keys to chests so that there is an abundance of chests compared to keys. Therefore, the environment may look like this instead:



Context change

The hope is that while training, the agent picks up a simple pseudo objective: collect as many keys as possible. Since chests are abundant, it shouldn't need to expend much energy seeking them, as it will nearly always run into one while traveling to the next key. Note that we can limit the number of steps during a training episode so that it almost never runs out of keys during training.

When taken off the training distribution, we can run this scenario in reverse. Instead of testing it in an environment with few keys and lots of chests, we can test it in an environment with few chests and many keys. Therefore, when pursuing the pseudo objective, it will spend all its time collecting keys without getting any reward.

Testing for mesa misalignment

In order to show that the mesa optimizer is [competent but misaligned](#) we can put the agent in a maze-like environment much larger than any it was trained for. Then, we can provide it an abundance of keys relative to chests. If it can navigate the large maze and collect many keys comfortably while nonetheless opening few or no chests, then it has experienced a malign failure.

We can make this evidence for pseudo alignment even stronger by comparing the trained agent to two that we hard-code: one agent that pursues the optimal policy for collecting keys, and one agent that pursues the optimal policy for opening as many chests as possible. Qualitatively, if the trained agent is more similar to the first agent than the second, then we should be confident that it has picked up the pseudo objective.

The analogy with human evolution

In the ancestral environment, calories were scarce. In our modern day world they are no longer scarce, yet we still crave them, sometimes to the point where it harms our reproductive capability. This is similar to how the agent will continue pursuing keys even if it is not using them to open any chests.

AI Safety "Success Stories"

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

AI safety researchers often describe their long term goals as building "safe and efficient AIs", but don't always mean the same thing by this or other seemingly similar phrases. Asking about their "success stories" (i.e., scenarios in which their line of research helps contribute to a positive outcome) can help make clear what their actual research aims are. Knowing such scenarios also makes it easier to compare the ambition, difficulty, and other attributes of different lines of AI safety research. I hope this contributes to improved communication and coordination between different groups of people working on AI risk.

In the rest of the post, I describe some common AI safety success stories that I've heard over the years and then compare them along a number of dimensions. They are listed in roughly the order in which they first came to my attention. (Suggestions welcome for better names for any of these scenarios, as well as additional success stories and additional dimensions along which they can be compared.)

The Success Stories

Sovereign Singleton

AKA [Friendly AI](#), an autonomous, superhumanly intelligent AGI that takes over the world and optimizes it according to some (perhaps indirect) specification of human values.

Pivotal Tool

An oracle or [task AGI](#), which can be used to perform a [pivotal](#) but limited act, and then stops to wait for further instructions.

Corrigible Contender

A semi-autonomous AGI that does not have long-term preferences of its own but acts according to (its understanding of) the [short-term preferences](#) of some human or group of humans, it competes effectively with comparable AGIs corrigible to other users as well as unaligned AGIs (if any exist), for resources and ultimately for influence on the future of the universe.

Interim Quality-of-Life Improver

AI risk can be minimized if world powers coordinate to limit AI capabilities development or deployment, in order to give AI safety researchers more time to figure out how to build a very safe and highly capable AGI. While that is proceeding, it may

be a good idea (e.g., politically advisable and/or morally correct) to deploy relatively safe, limited AIs that can improve people's quality of life but are not necessarily state of the art in terms of capability or efficiency. Such improvements can for example include curing diseases and solving pressing scientific and technological problems.

(I want to credit Rohin Shah as the person that I got this success story from, but can't find the post or comment where he talked about it. Was it someone else?)

Research Assistant

If an AGI project gains a lead over its competitors, it may be able to grow that into a larger lead by building AIs to help with (either safety or capability) research. This can be in the form of an oracle, or human imitation, or even narrow AIs useful for making money (which can be used to buy more compute, hire more human researchers, etc). Such Research Assistant AIs can help pave the way to one of the other, more definitive success stories. Examples: [1](#), [2](#).

Comparison Table

| | Sovereign Singleton | Pivotal Tool | Corrigible Contender | Interim Quality-of-Life Improver | Research Assistant |
|--|---------------------|--------------|----------------------|----------------------------------|--------------------|
| Autonomy | High | Low | Medium | Low | Low |
| AI safety ambition / difficulty | Very High | Medium | High | Low | Low |
| Reliance on human safety | Low | High | High | Medium | Medium |
| Required capability advantage over competing agents | High | High | None | None | Low |
| Tolerates capability trade-off due to safety measures | Yes | Yes | No | Yes | Some |
| Assumes strong global coordination | No | No | No | Yes | No |
| Controlled access | Yes | Yes | No | Yes | Yes |

(Note that due to limited space, I've left out a couple of scenarios which are straightforward recombinations of the above success stories, namely Sovereign Contender and Corrigible Singleton. I also left out CAIS because I find it hard to visualize it clearly enough as a success story to fill out its entries in the above table, plus I'm not sure if any safety researchers are currently aiming for it as a success story.)

The color coding in the table indicates how hard it would be to achieve the required condition for a success story to come to pass, with green meaning relatively easy, and yellow/pink/violet indicating increasing difficulty. Below is an explanation of what each row heading means, in case it's not immediately clear.

Autonomy

The opposite of human-in-the-loop.

AI safety ambition/difficulty

Achieving each success story requires solving a different set of AI safety problems. This is my subjective estimate of how ambitious/difficult the corresponding set of AI safety problems is. (Please feel free to disagree in the comments!)

Reliance on human safety

How much does achieving this success story depend on humans being safe, or on solving human safety problems? This is also a subjective judgement because different success stories rely on different aspects of human safety.

Required capability advantage over competing agents

Does achieving this success story require that the safe/aligned AI have a capability advantage over other agents in the world?

Tolerates capability trade-off due to safety measures

Many ways of achieving AI safety have a cost in terms of lowering the capability of an AI relative to an unaligned AI built using comparable resources and technology. In some scenarios this is not as consequential (e.g., because it depends on achieving a large initial capability lead and then preventing any subsequent competitors from arising), and that's indicated by a "Yes" in this row.

Assumes strong global coordination

Does this success story assume that there is strong global coordination to prevent unaligned competitors from arising?

Controlled access

Does this success story assume that only a small number of people are given access to the safe/aligned AI?

Further Thoughts

1. This exercise made me realize that I'm confused about how the Pivotal Tool scenario is supposed to work, after the initial pivotal act is done. It would likely require several years or decades to fully solve AI safety/alignment and remove the dependence on human safety, but it's not clear how to create a safe environment for doing that after the pivotal act.
2. One thing I'm less confused about now is why people who work toward the Contender scenarios are focused more on minimizing the capability trade-off of safety measures than people who work toward the Singleton scenarios even though the latter scenarios seem to demand more of a capability lead. It's because the latter group of people think it's possible or likely for a single AGI project to achieve a large initial capability advantage, in which case some initial capability trade-off due to safety measures is ok, and subsequent ongoing capability trade-off is not consequential because there would be no competitors left.
3. The comparison table makes Research Assistant seem a particularly attractive scenario to aim for, as a stepping stone to a more definitive success story. Is this conclusion actually justified?
4. Interim Quality-of-Life Improver also looks very attractive, if only strong global coordination could be achieved.

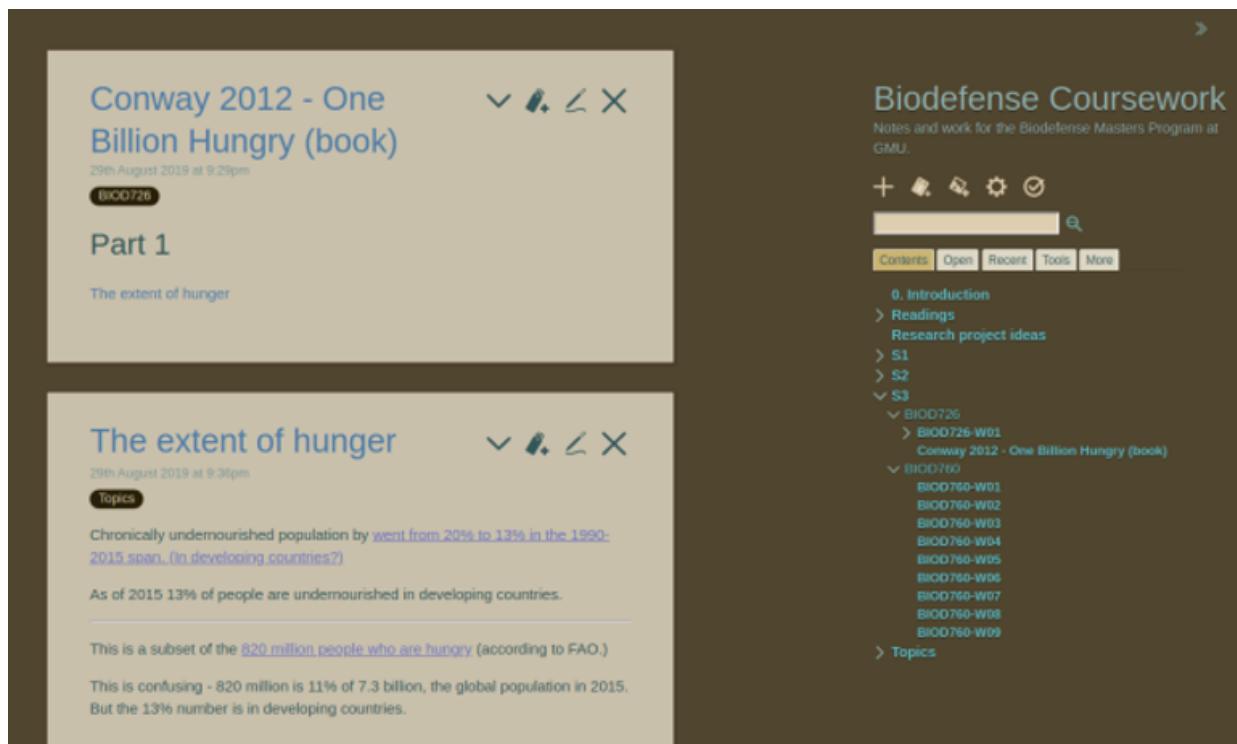
Tiddlywiki for organizing notes and research

This is a linkpost for <https://eukaryotewritesblog.com/2019/09/01/tiddlywiki-for-organizing-notes-and-research/>

Happy new school year to my fellow students! With my first year of grad school under my belt, and my sword and shield out for Round 2, I wanted to share a tool that's helped me on my journey.

Two years ago, my go-to system for organizing my research and writing "citations in 3 different programs" + "pile everything into a haphazard series of google docs and hope for the best". I figured this wasn't great. After doing some reading and trying several alternatives, I discovered Tiddlywiki.

[Tiddlywiki](#) is an ancient open-source wiki application in the form of an html file. It has all the tools you need to make a wiki in the form of "tiddlers", self-contained chunks of info that you can tag and link to each other. When you save the wiki, the program and your text all get wrapped up together into the same .html file - it both stores your info and is the program for running the wiki. It works on any web browser, as well as special programs.



I stuck with it and here's why:

- **Wiki format:** Wikis seem really compatible with the way my brain works. If I take notes on a book or article, that source gets its own tiddler on the wiki. They can then get interwoven, crosslinked, expanded upon, etc.

- **Elegant:** Does most things I want it to. Easy to link to tiddlers and drag them or other files in from other wikis/folders. The structure is transparent and customizable.
- **Robust:** Tiddlywikis from a decade ago are still perfectly functional today. The entire program and dataset lives in one small html file that runs on anything with a web browser.
- **Meta-aesthetics:** Feeding all my data to Google is a little worrying. Tiddlywiki, meanwhile, is open-source and runs from your computer. The fact that the program is a quine is really neat.
- **Encryption:** Tiddlywikis have an encryption function baked in. I don't know if it's very good. Consider using [Veracrypt](#) for better security. But if you don't want to do that, here you go. This also means you can upload your wikis and backups to cloud services while keeping them encrypted. (Go to "Tools" in the sidebar, then click on the "set password" button. After you set a password, you can look at the .html file text to be sure that, yes, everything is encrypted into nonsense characters.)
- **Customizable:** Easily change the color scheme, any text or formatting, the layout, etc. It's extremely adaptable. You can also install a variety of plugins, though I haven't felt the need to myself as of yet.
- **Transportable:** My wikis live on a flashdrive and can work on any computer. I took all my research with me to and from work every day this summer for an internship.

Things I like less

- **Saving** is not obvious. This simplest version is "edit a copy of a blank tiddlywiki in a web browser, save locally to your computer or a flash drive, repeat every time you edit it", which is kind of a pain. ** I work on various computers, so my Tiddlywikis are saved on a flash drive. I edit them in web browsers, and save them back to the flash drive when I'm done. I back them up every week. ** On my Ubuntu laptop, I edit them with the program [TiddlyDesktop](#), which makes saving easier.
- You can use **images**, but they get saved as raw code into the html file itself (so every image makes the file that much larger), and there aren't tools for manipulating them. (There is a cute, tiny, and almost useless drawing program baked in.) I tend to save a few images, like graphs or figures from papers, but wouldn't personally use Tiddlywiki for image-heavy work.
- Some features (e.g. spellcheck, in-text search with highlighting) **depend on the browser** or other program you're using to edit the wikis.
- Kind of old-looking, **not maximally aesthetic**.

The number of wikis you have is up to you. I started with one wiki for a specific writing project and one wiki for work, notes and research. My active Tiddlywikis now include:

- Grad school material
- Internship research material
- General writing, notes, and personal research
- Writing and worldbuilding/characterization/plot details for a novel
- Recipe storage
- Quotes and poetry I like

Your mileage may vary.

How do I try it?

First, check out some tiddlywikis that have been converted into websites. [Here's a nice one to explore as an example](#), a thesis website in Spanish. [Here's one on philosophy](#). (Note that you can't actually edit the versions that appear on the website. You can locally save the whole wiki and changes you make to it, though.)

If you like it, here are some resources to get you started. [This is the official website](#), which has lots of helpful documentation. (Note that it's also a tiddlywiki!)

[Here are some youtube videos](#) I also found helpful.

After making a few tiddlywikis, I found that I kept making the same tweaks to them to get them set up in a way useful for me. In that light, I made a new “blank” or “empty” tiddlywiki that had those changes baked in already.

Here it is: [the Eukaryote Writes Blog empty tiddlywiki](#). You may find it better than the default empty wiki. It comes with a couple new color schemes, a table of contents, and some layout tweaks, among other small changes.

Other research tools

All hail the [exobrain](#)!

I keep track of research citations formally with [Zotero](#), or the tool my work prefers. For informal reading, I'll also just note the authors and title and/or URL of the source (in my tiddlywiki!) so I can find it later.

For keeping track of time spent working, I've gotten some utility out of [KanbanFlow](#). I like [the Pomodoro Technique](#), and KanbanFlow has both pomodoro timers and a nice task-sorting and task-prioritization system built in. I currently don't worry about tracking time, and use Google Calendar, [a bullet journal](#), and a [bastardized Kanban Board variant](#) to keep my brain in order.

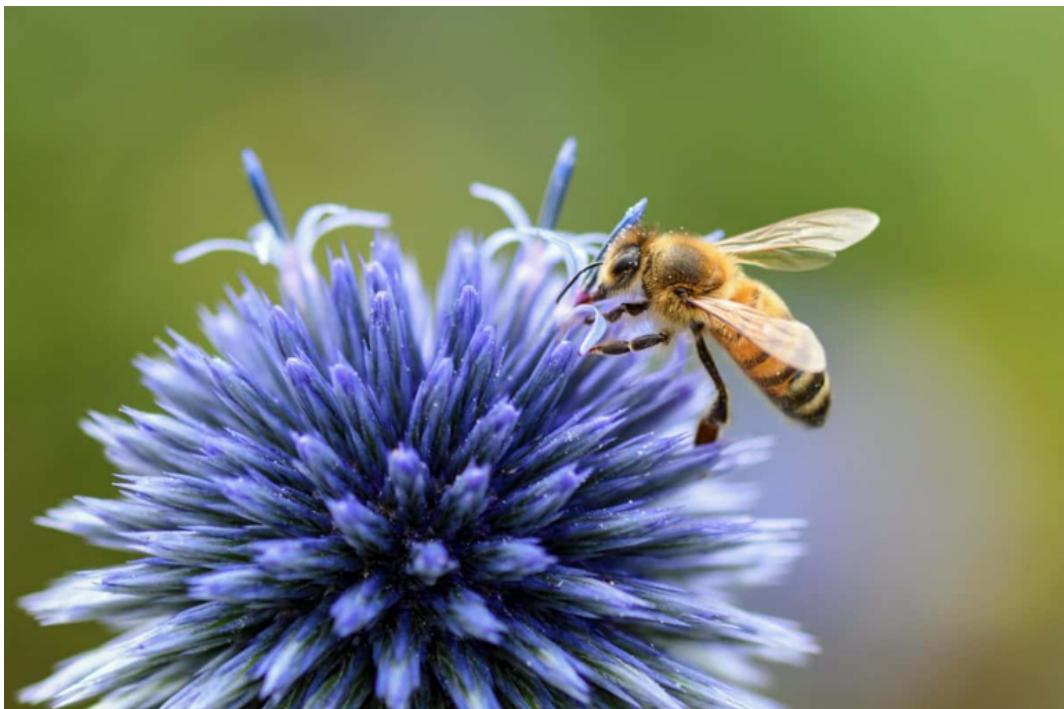
Previously, I used the website [MarinaraTimer](#) to time pomodoros. I love it for exactly two reasons: the ability to pause pomodoros, and the sound effect “Ominous Woosh”.

Reframing the evolutionary benefit of sex

Automatically crossposted from <https://sideways-view.com>

From the perspective of an organism trying to propagate its genes, sex is like a trade: I'll put half of your DNA in my offspring if you put half of my DNA in yours. I still pass one copy of my genes onto the next generation per unit of investment in children, so it's a fair deal. And it doesn't impact the average fitness of my kids very much, since on average my partner's genes will be about as good as mine. (ETA: but see the discussion below, in which case the costs might be much bigger.)

But the trade has transaction costs, so I'm only going to do it if I get some benefit. In this post I'll tell a particularly simple story about the benefit of sex. I think this is basically equivalent to the [standard story](#), but I find it much clearer. It also makes it more obvious that we don't require group selection, and that the benefit is very large.



Why doesn't sex change the average fitness of my kids? The possibility of a "lucky" kid who gets the better genes from both of us is offset by the possibility of an unlucky kid who gets the worse genes from both of us. If the effects of genes are linear, the average fitness will be exactly the same as the parents. In practice I expect it to be slightly lower because of convexity and linkage disequilibrium.

But sex increases the average fitness of my *grandchildren*, because my fittest children will be responsible for a disproportionate fraction of my grandkids. More precisely, if my if an organism with fitness dX has $(1+dX)$ kids per generation, then the total fitness of my grandkids is $E[(1 + dX)^2] = 1 + 2 E[dX] + E[dX^2]$. So increasing variance by 1 unit is as good as increasing average fitness by 0.5 units.

Reproductive decisions are naturally a tradeoff between average fitness and variance. Sex slightly lowers the average but increases the variance. If you try to get the same amount of variance with random mutations, you'll have to totally tank your kid's expected fitness, because your current genome is well-optimized, **and** you'll also pass on fewer genes to the next generation (since some of yours got destroyed). In fact, it's hard to think of any way to get similar benefits without exchanging genes.

Variance becomes linearly more important over time. For the fitness of my grandkids, 1 unit of variance is worth 0.5 units of fitness; for great-grandkids, they are equally valuable; for great-great-great-grandkids variance is twice as valuable. I don't think we have to look too many generations ahead before the variance bonus from sex outweighs the costs. For example, if genetic fitness differs by 5% between my offspring, and sex reduces fitness by 1%, then sex breaks even within 6 generations.

Free Money at PredictIt?

Previously on Prediction Markets (among others): [Prediction Markets: When Do They Work?, Subsidizing Prediction Markets](#)

Epistemic Status: No huge new insights, but a little fun, a little free money, also [Happy Petrov Day?](#)

Yesterday, with everything happening regarding impeachment, I decided to check PredictIt to find out how impactful things were. When I checked, I noticed some obvious inconsistencies. They're slightly less bad today, but still not gone.

I figured it would be fun and potentially enlightening to break this down. Before I begin, I will state that unless I messed up this post expresses *zero political opinions whatsoever* on what election or other outcomes would be good or bad, and does its best to only make what I consider *very safe observations* on probabilities of events. All comments advocating political positions or candidates will be deleted in reign-of-terror style. No exceptions.

Market Analysis

Odds are represented as cost in cents for each contract that pays \$1, so they double as probabilities out of 100.

Let us look at the democratic nomination odds, using last. All are 1 cent wide:

Elizabeth Warren 50

Joe Biden 21

Andrew Yang 10

Bernie Sanders 8

Pete Buttigieg 6

Hillary Clinton 5

Kamala Harris 4

Tulsi Gabbard 3

Amy Klobuchar 2

Can be sold for 1: Corey Booker, Tom Snyder, Beto 'o Rourke

All other candidates can be bought for 1 and cannot be sold.

Adding that up we get 112. We could buy all the no sides for a total of 111 - you can get these prices on no except for Warren, where you'd sell at 49.

That's certainly some free money. If you sell all of them, you don't tie up any money, although you do have to deposit, so it's a pretty great trade, albeit with an \$850 limit.

I've already done many of the legs of that trade. Some are better than others. Hillary Clinton at 5 is complete insanity. Andrew Yang is trading at 10 because internet. That likely covers most of the reason you can sell the field for 111. Lower them to sane numbers (let's be super generous and say Hillary Clinton 1, and say Andrew Yang 5) then the field would add to 102. Completing the trade is mostly about freeing up your capital. You also get *some* value for it being someone not on the above list, as the 'brokered convention causes weirdness' scenario is definitely not impossible. The weird thing is expecting that to somehow nominate Hillary Clinton.

The big not-automatically-insane opinion is making Warren 50% to win the nomination. That is rather bold at this stage of things, but we're thinking about arbitrage and outright mistakes.

Let's now look at the Presidential odds. For any Democrat, this is almost identical to a two-part bet, where that person wins the nomination and then wins the general election.

Donald Trump 41

Elizabeth Warren 35

Joe Biden 13

Andrew Yang 6

Bernie Sanders 6

Pete Buttigieg 3

Nikki Haley 2

Kamala Harris 2

Mike Pence 2

Tulsi Gabbard 2

Corey Booker 1

Amy Klobuchar 1

That adds up to 114. If you look at actually available prices, you could sell the field for 110. Again, pretty good idea. I'd get on that, and I mostly did.

One could also point out the implied general election win percentages of democrats where rounding isn't a big deal.

Warren 70%

Biden 62%

Yang 60%

Sanders 75%

Sum of All Republican odds is 45% (Trump, Haley and Pence) out of 114%, for odds of 39.4%. Thus, Democratic victory should be about 60%. Warren is 50% to win the nomination, so that 70% number is really weird. This does not add up, and makes me reluctant to sell Warren at 50% odds in the primary.

In both these cases, the free money seems real enough. You get to use your capital in both markets if you sell the whole field and then have it free for a third market as well, and you can't really lose. Doesn't mean it's worth the effort, but it's a nice thing to notice.

Let's look at Republican nomination odds:

| | |
|---------|----|
| Trump | 78 |
| Haley | 7 |
| Pence | 7 |
| Kasich | 2 |
| Romney | 2 |
| Weld | 2 |
| Sanford | 2 |

That only adds up to 98%, which makes sense, since if Trump is actually gone then anything could happen. This market seems sane on that level, perhaps even rich. What's most interesting is that Trump is highly unlikely to not win the Republican nomination and win the presidency, so if he's 41% to be reelected but 78% to be nominated, then Trump has a general election win rate of 52% (47% if we knock off 10% for the market being inflated by adding up to 114%). But perhaps this is reasonable? If Trump is gone it's because something brought him down so it's going to be super hard for anyone else to win? It's not like much of that probability is that Trump's health fails, given the time frame.

Also noteworthy is Trump is only 20% to win *the popular vote*, although the available volume here is very low. That implies a stunning 21% chance that Trump loses the popular vote but wins the election. Put another way, given Trump is reelected, he's still an underdog to have won the popular vote. The electoral college seems to favor Trump, but that's a *huge* probability to put in such a narrow space, even if you assume the states all look identical to 2016. I believe that pre-election, 538 had Trump at 10% to do this, with the polls only a few percent away from that result. How do you get to 20%?

You can sell "Hillary Clinton runs for president in 2020" at 12% odds. Is that a worthwhile return on capital? You could also sell Michele Obama at 8%, Cuomo at 6%, and Oprah or Mark Cuban at 5%.

They have Trump at 88% to be President at the end of 2019 and 73% to complete his first term. They think he is 41% to be impeached this year and 63% to be impeached at all. Congress is expected to work fast. Have they met congress?

There are a number of other similar good bets available. One gets the idea. The catch is that those all tie up capital. Also, if you take risk and win, you have to pay 10% of

your net winnings and potentially taxes. Again, three cheers for arbitrage.

Looking at such systems of prices, and looking for opportunity, is often good training as not only a trader or gambler but also for calibration and probability estimation in general, which are excellent skills for anyone to develop.

What Does This Say About Prediction Markets in General?

Not much we didn't already know. PredictIt has an \$850 limit on any one market, for any one candidate or other potential outcome. This does not increase if you do arbitrage. This is why pure arbitrage that frees up capital can continue. I am literally at risk for \$44 in the general election market, but that does not allow me to continue to trade.

Other markets in the past such as InTrade have not had this restriction. This results in less egregious versions of the same problems, as you can use bigger size to trade against the mistakes. However, there is no point in *fully correcting* a mistake, as doing so would offer minimal or no profits. If you have a market that is inefficient, and a chance to trade to make it more efficient, that's a good trade, but at some point it isn't worth the time and trouble and capital investment, so you stop. That point is *necessarily* before full efficiency, but in places like the stock market you can potentially get (in expectation) very close.

In prediction markets, cost of capital to do trades is a major distorting factor, as are fees and taxes and other physical costs, and participants are much less certain of correct prices and much more worried about impact and how many others are in the same trade. Most everyone who is looking to correct inefficiencies will only fade *very large and very obvious* inefficiencies, given all the costs.

Thus, we see the same inefficiencies pop up over and over again and not be corrected. The most well-known and universal one is that if the probability is under about 40%, the odds will likely be too high. The lower the odds below that, the more (as a percentage of the chance listed) the price will be too high. For low percentages, the people selling the contract are treating it as if it is a bond that pays interest over time, with a tiny default risk, rather than saying that the 7% chance is too high and should have been 5%. One also has to be wary of model error.

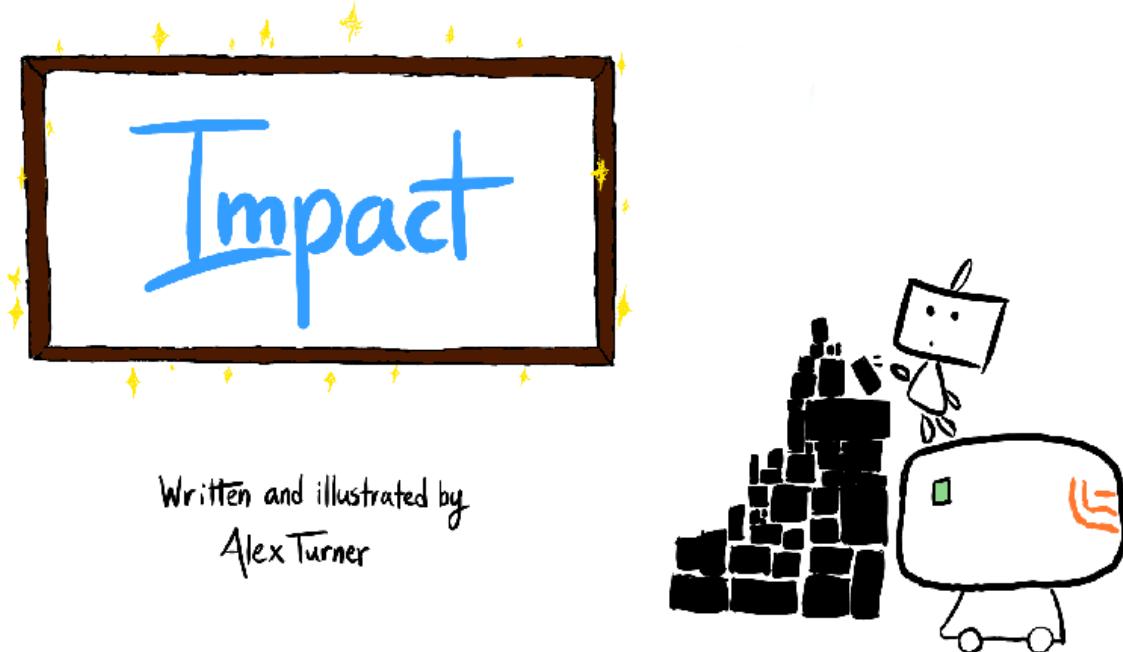
In politics, it is also inevitable that anything that sounds superficially good to people on the internet but is unlikely to actually happen is almost always going to trade rich.

If anything, it is remarkable *how little difference* it made to limit accounts to \$850 in trading, beyond there being free cash lying around.

Anyway, thought that would be fun to write up formally given I had been tricked into actually trading the markets, and maybe some of you would get to do some good trades, so I figured why not. Have fun, everyone.

Reframing Impact

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.



Imagine we have a robot named Frank.

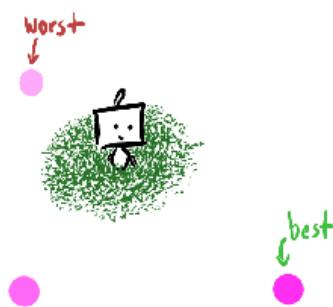
Frank finds things for us in places we can't go.

We provide a rule, and he returns with the object that best fits the rule.

Right now, we **want** a very pink marble.

(in case you're wondering)

Naturally, we ask Frank for the pinkest thing he can find.

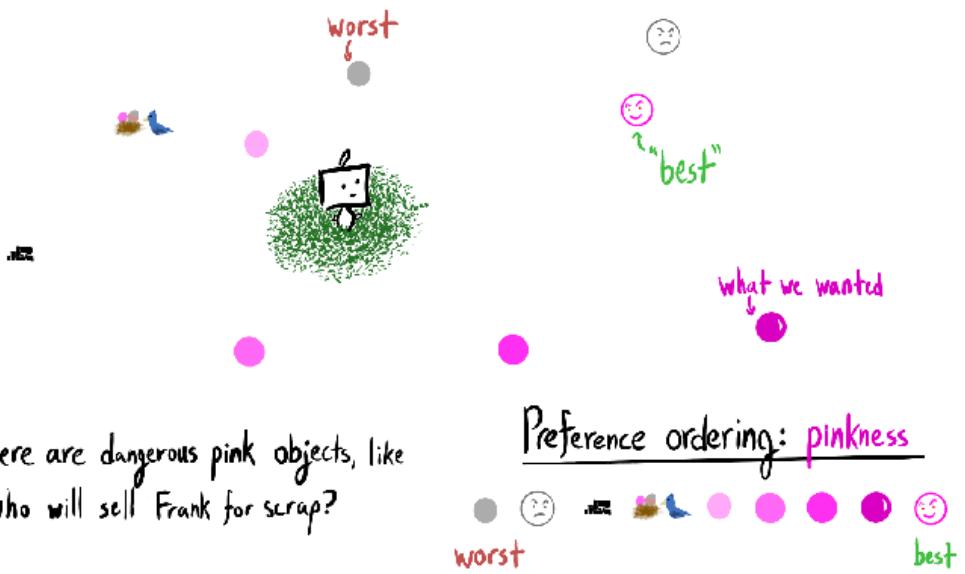


Preference ordering: pinkness

● worst ● best

This seems fine. But what if Frank looks farther afield?

The world is wide, and full of objects.



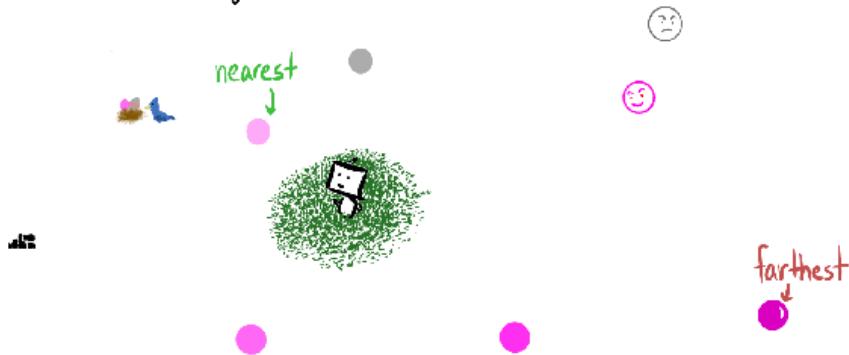
What if there are dangerous pink objects, like
terrorists who will sell Frank for scrap?

Preference ordering: pinkness

● ☹ - ■ ☺
worst best

From our perspective, Frank has lost his marbles, but he's just following an imperfect rule.
The perfect rule is hard to specify. What simple rule avoids terrorists here?

Fortunately, the terrorists are far away.



Pinkness correlates with what we want,
and the proximity rule avoids terrorists.

Preference ordering: proximity



Think about what Frank brings us for each distance.



We're probably fine with a reasonably pink marble.
Then how about we have Frank find the pinkest object
within a given distance, which we increase until we're satisfied?

And now for the reveal

Frank is analogous to a powerful AI with an imperfect **objective**. The objects are plans he's considering, and the terrorists are catastrophic plans (some of which happen to **score well**).

The question is then:



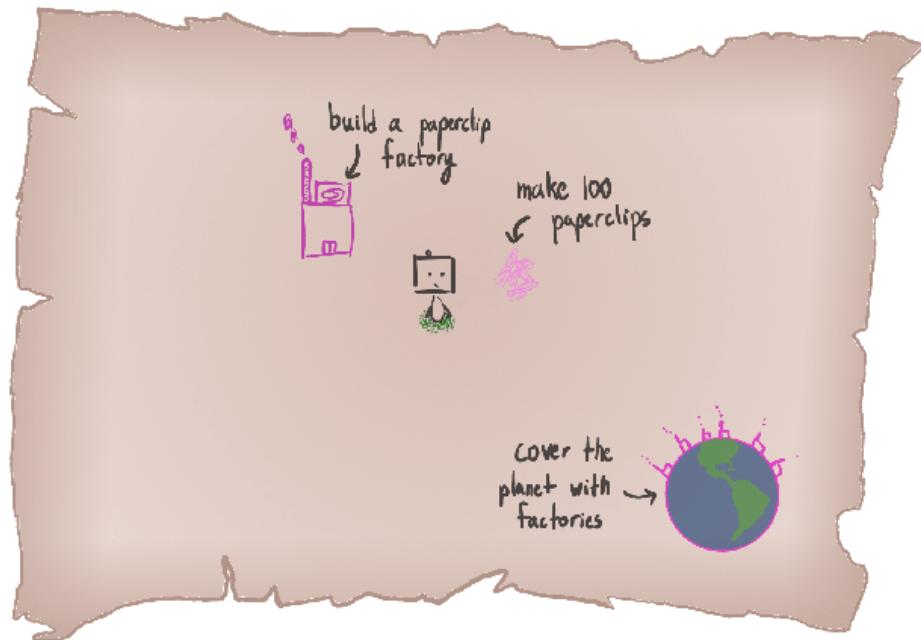
How do we measure how
distant
plans are?

The *distance measure* should:

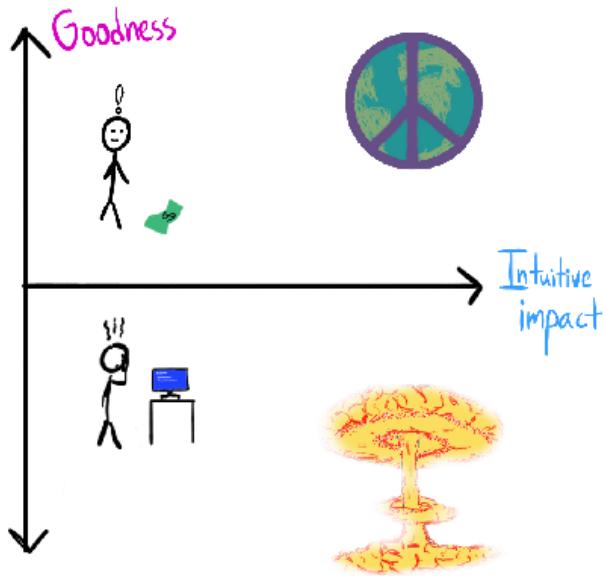
- 1) Be easy to specify
- 2) Put catastrophes far away
- 3) Put reasonable plans nearby



Suppose Frank's *objective* is to make paperclips.
The *measure* will put plans on the map:



These catastrophes seem like **big deals**. We're going to figure out **why** we intuit some things are **big deals**, develop an understanding of the relevant parts of reality, and then design an **impact measure**.



To me, the **impactful** things feel fundamentally different than the non-**impactful** things. I find this difference fascinating and beautiful, and look forward to exploring it with you.

Why be excited about **impact measurement**? After all, it doesn't seem like a one-shot solution to AI alignment. It doesn't even seem like a key problem, like figuring out how to get AIs to robustly learn **human values**, or understanding what it means to be both an **agent and part** of the environment.

Frankly, this misses the forest for the trees. At its core, **impact measurement** is about how, when, and why agents **affect** each other. Understanding this requires a new way of looking at the world.

The question of **impact measurement** has caused significant confusion, but together, we'll find comprehension. We're going to emerge **saddled with spoils**: new conceptual frameworks, fresh lines of inquiry, and important theoretical milestones.

Here's one exciting milestone we're shooting for:



An impact measure would be the first proposed safeguard which maybe actually stops a powerful agent with an imperfect objective from ruining things – without assuming anything about the objective. This is a rare property among approaches.



We have our bearing.
Let us set out together



Technical Appendix: First safeguard?

This sequence is written to be broadly accessible, although perhaps its focus on capable AI systems assumes familiarity with [basic arguments for the importance of AI alignment](#). The technical appendices are an exception, targeting the technically inclined.

Why do I claim that an impact measure would be "the first proposed safeguard which maybe actually stops a powerful agent with an imperfect objective from ruining things – without assuming anything about the objective"?

The safeguard proposal shouldn't have to say "and here we solve this opaque, hard problem, and then it works". If we have the impact measure, we have the math, and then we have the code.

So what about:

- Quantilizers? This seems to be the most plausible alternative; mild optimization and impact measurement share many properties. But

- What happens if the agent is already powerful? A greater proportion of plans could be catastrophic, since the agent is in a better position to cause them.
- Where does the base distribution come from (opaque, hard problem?), and how do we know it's safe to sample from?
 - In the linked paper, Jessica Taylor suggests the idea of learning a human distribution over actions – how robustly would we need to learn this distribution? How numerous are catastrophic plans, and what *is* a catastrophe, defined without reference to our values in particular? (That definition requires understanding impact!)
- [Value learning](#)? But
 - We only want this if *our* (human) values are learned!
 - [Value learning is impossible without assumptions](#), and [getting good enough assumptions could be really hard](#). If we don't know if we can get value learning / reward specification right, we'd like safeguards which don't fail because value learning goes wrong. The point of a safeguard is that it can catch you if the main thing falls through; if the safeguard fails because the main thing does, that's pointless.
- [Corrigibility](#)? At present, I'm excited about this property because I suspect it has a simple core principle. But
 - Even if the system is responsive to correction (and non-manipulative, and whatever other properties we associate with corrigibility), what if we become *unable* to correct it as a result of early actions (if the agent "moves too quickly", so to speak)?
 - [Paul Christiano's take on corrigibility](#) is much broader and an exception to this critique.
 - What is the core principle?

Notes

- The three sections of this sequence will respectively answer three questions:
 - Why do we think some things are big deals?
 - Why are capable goal-directed AIs incentivized to catastrophically affect us by default?
 - How might we build agents without these incentives?
- The first part of this sequence focuses on foundational concepts crucial for understanding the deeper nature of impact. We will not yet be discussing what to implement.
- I strongly encourage completing the exercises. At times you shall be given a time limit; it's important to learn not only to reason correctly, but with speed.

The best way to use this book is NOT to simply read it or study it, but to read a question and STOP. Even close the book. Even put it away and THINK about the question. Only after you have formed a reasoned opinion should you read the solution. Why torture yourself thinking? Why jog? Why do push-ups?

If you are given a hammer with which to drive nails at the age of three you may think to yourself, "OK, nice." But if you are given a hard rock with which to drive nails at the age of three, and at the age of four you are given a hammer, you think to yourself, "What a marvellous invention!" You see, you can't really appreciate the solution until you first appreciate the problem.

~ [Thinking Physics](#)

- My paperclip-Balrog illustration is metaphorical: a good impact measure would hold steadfast against the daunting challenge of formally asking for the right thing from a powerful agent. The illustration does not represent an internal conflict within that agent. As water flows downhill, an impact-penalizing Frank prefers low-impact plans.
 - The drawing is based on [gonzalokenny's amazing work](#).

- Some of you may have a different conception of impact; I ask that you grasp the thing that I'm pointing to. In doing so, you might come to see your mental algorithm is the same. Ask not “is this what I initially had in mind?”, but rather “does this make sense as a thing-to-call-'impact'?”.
- H/T Rohin Shah for suggesting the three key properties. Alison Bowden contributed several small drawings and enormous help with earlier drafts.

The Power to Demolish Bad Arguments

This is Part I of the [Specificity Sequence](#)

Imagine you've played ordinary chess your whole life, until one day the game becomes 3D. That's what unlocking the power of specificity feels like: a new dimension you suddenly perceive all concepts to have. By learning to navigate the specificity dimension, you'll be training a unique mental superpower. With it, you'll be able to jump outside the ordinary course of arguments and fly through the conceptual landscape. Fly, I say!



"Acme exploits its workers!"

Want to see what a 3D argument looks like? Consider a conversation I had the other day when my friend "Steve" put forward a claim that seemed counter to my own worldview:

Steve: Acme exploits its workers by paying them too little!

We were only one sentence into the conversation and my understanding of Steve's point was high-level, devoid of specific detail. But I assumed that whatever his exact point was, I could refute it using my understanding of basic economics. So I shot back with a counterpoint:

Liron: No, job creation is a force for good at any wage. Acme increases the demand for labor, which drives wages up in the economy as a whole.

I injected principles of Econ 101 into the discussion because I figured they could help me expose that Steve misunderstood Acme's impact on its workers.

My smart-sounding response might let me pass for an intelligent conversationalist in non-rationalist circles. But my rationalist friends wouldn't have been impressed at my 2D tactics, parrying Steve's point with my own counterpoint. They'd have sensed that I'm not progressing toward clarity and mutual understanding, that I'm ignorant of the way of [Double Crux](#).

If I were playing 3D chess, my opening move wouldn't be to slide a defensive piece (the Econ 101 Rook) across the board. It would be to... shove my face at the board and stare at the pieces from an inch away.

Here's what an attempt to do this might look like:

Steve: Acme exploits its workers by paying them too little!

Liron: What do you mean by "exploits its workers"?

Steve: Come on, you know what "exploit" means... Dictionary.com says it means "to use selfishly for one's own ends"

Liron: You're saying you have a beef with any company that acts "selfish"? Doesn't every company under capitalism aim to maximize returns for its shareholders?

Steve: Capitalism can be good sometimes, but Acme has gone beyond the pale with their exploitation of workers. They're basically ruining capitalism.

No, this is still not the enlightening conversation we were hoping for.

But where did I go wrong? Wasn't I making a respectable attempt to lead the conversation toward clear and precise definitions? Wasn't I navigating the first waypoint on the road to Double Crux?

Can you figure out where I went wrong?

...

...

...

It was a mistake for me to ask Steve for a mere *definition* of the term "exploit". I should have asked for a *specific example* of what he imagines "exploit" to mean in this context. What specifically does it mean—actually, forget "mean"—what specifically does it *look like* for Acme to "exploit its workers by paying them too little"?

When Steve explained that "exploit" means "to use selfishly", he fulfilled my request for a definition, but the whole back-and-forth didn't yield any insight for either of us. In retrospect, it was a wasted motion for me to ask, "What do you mean by 'exploits its workers'".

Then I, instead of making another attempt to shove my face to stare at the board up close, couldn't help myself: I went back to sliding my pieces around. I set out to rebut the claim that "Acme uses its workers selfishly" by tossing the big abstract concept of "capitalism" into the discussion.

At this point, imagine that Steve were a malicious actor whose only goal was to score rhetorical points on me. He'd be thrilled to hear me say the word "capitalism". "Capitalism" is a nice high-level concept for him to build an infinite variety of smart-sounding defenses out of, together with other high-level concepts like "exploitation" and "selfishness".

A malicious Steve can be rhetorically effective against me even without possessing a structured understanding of the subject he's making a claim about. His mental model of the

subject could just be a ball pit of loosely-associated concepts. He can hold up his end of the conversation surprisingly well by just snatching a nearby ball and flinging it at me. And what have I done by mentioning “capitalism”? I’ve gone and tossed in another ball.



I'd like to think that Steve isn't malicious, that he isn't trying to score rhetorical points on me, and that the point he's trying to make has some structure and depth to it. But there's only one way to be sure: By using the power of specificity to get a closer look! Here's how it's done:

Steve: Acme exploits its workers by paying them too little!

Liron: Can you help me paint a specific mental picture of a worker being exploited by Acme?

Steve: Ok... A single dad who works at Acme and never gets to spend time with his kids because he works so much. He's living paycheck to paycheck and he doesn't get any paid vacation days. The next time his car breaks down, he won't even be able to fix it because he barely makes minimum wage. You should try living on minimum wage so you can see how hard it is!

Liron: You're saying Acme should be blamed for this specific person's unpleasant life circumstances, right?

Steve: Yes, because they have thousands of workers in these kinds of circumstances, and meanwhile their stock is worth \$80 billion.

Steve doesn't realize this yet, but by coaxing out a specific example of his claim, I've suddenly made it impossible for him to use a ball pit of loosely-associated concepts to score rhetorical points on me. From this point on, the only way he can win the argument with me is by clarifying and supporting his claim in a way that helps me update my mental model of the subject. This isn't your average 2D argument anymore. We're now flying like Superman.



Liron: Ok, sticking with this one specific worker's hypothetical story—what would they be doing if Acme didn't exist?

Steve: Getting a different job

Liron: Ok, what specific job?

Steve: I don't know, depends what their skills are

Liron: This is your specific story Steve, you get to pick any specific plausible details you want in order to support any point you want!

I have to stop and point out how crazy this is.

You'd think the way smart people argue is by supporting their claims with evidence, right? But here I'm giving Steve a handicap where he gets to *make up fake evidence* (telling me any hypothetical specific story) just to establish that his argument is *coherent* by checking whether empirical support for it ever *could meaningfully exist*. I'm asking Steve to step over a really low bar here.

Surprisingly, in real-world arguments, this lowly bar often stops people in their tracks. The conversation often goes like this:

Steve: I guess he could instead be a cashier at McDonald's. Because then he'd at least get three weeks per year of paid vacation time.

Liron: In a world where Acme exists, couldn't this specific guy still go get a job as a cashier at McDonald's? Plus, wouldn't he have less competition for that McDonald's cashier job because some of the other would-be applicants got recruited to be Acme

workers instead? Can we conclude that the specific person who you chose to illustrate your point is actually being *helped* by the existence of Acme?

Steve: No because he's an Acme worker, not a McDonald's cashier

Liron: So doesn't that mean Acme offered him a better deal than McDonald's, thereby improving his life?

Steve: No, Acme just tricked him into thinking that it's a better deal because they run ads saying how flexible their job is, and how you can set your own hours. But it's actually a worse deal for the worker.

Liron: So like, McDonald's offered him \$13/hr plus three weeks per year of paid vacation, while Acme offered \$13/hr with no paid vacation time, but more flexibility to set his own hours?

Steve: Um, ya, something like that.

Liron: So if Acme did a better job of educating prospective workers that they don't offer the same paid vacation time that some other companies do, then would you stop saying that Acme is "exploiting its workers by paying them too little"?

Steve: No, because Acme preys on uneducated workers who need quick access to cash, and they also intend to automate away the workers' jobs as soon as they can.

Liron: It looks like you're now making new claims that weren't represented in the specific story you chose, right?

Steve: Yes, but I can tell other stories

Liron: But for the specific story you chose to tell that was supposed to best illustrate your claim, the "exploitation" you're referring to only deprived the worker of the value of a McDonald's cashier's paid vacation time, which might be like a 5% difference in total compensation value? And since his work schedule is much more flexible as an Acme worker, couldn't that easily be worth the 5% to him, so that he wasn't "tricked" into joining Acme but rather made a decision in rational self-interest?

Steve: Yeah maybe, but anyway that's just one story.

Liron: No worries, we can start over and talk about a specific story that you think *would* illustrate your main claim. Who knows, I might even end up agreeing with your claim once I understand it. It's just hard to understand what you're saying about Acme without having at least one example in mind, even if it's a hypothetical one.

Steve thinks for a little while...

Steve: I don't know all the exploitative shit Acme does ok? I just think Acme is a greedy company.

When someone makes a claim you (think you) disagree with, don't immediately start gaming out which 2D moves you'll counterargue with. Instead, start by drilling down in the specificity dimension: think through one or more specific scenarios to which their claim applies.

If you can't think of *any* specific scenarios to which their claim applies, maybe it's because there are none. Maybe the thinking behind their original claim is incoherent.

In complex topics such as politics and economics, the sad reality is that people who think they're *arguing* for a claim are often not even *making* a claim. In the above conversation, I never got to a point where I was trying to *refute Steve's argument*, I was just trying to get specific clarity on *what Steve's claim is*, and I never could. We weren't able to discuss an

example of what specific world-state constitutes, in his judgment, a referent of the statement “Acme exploits its workers by paying them too little”.

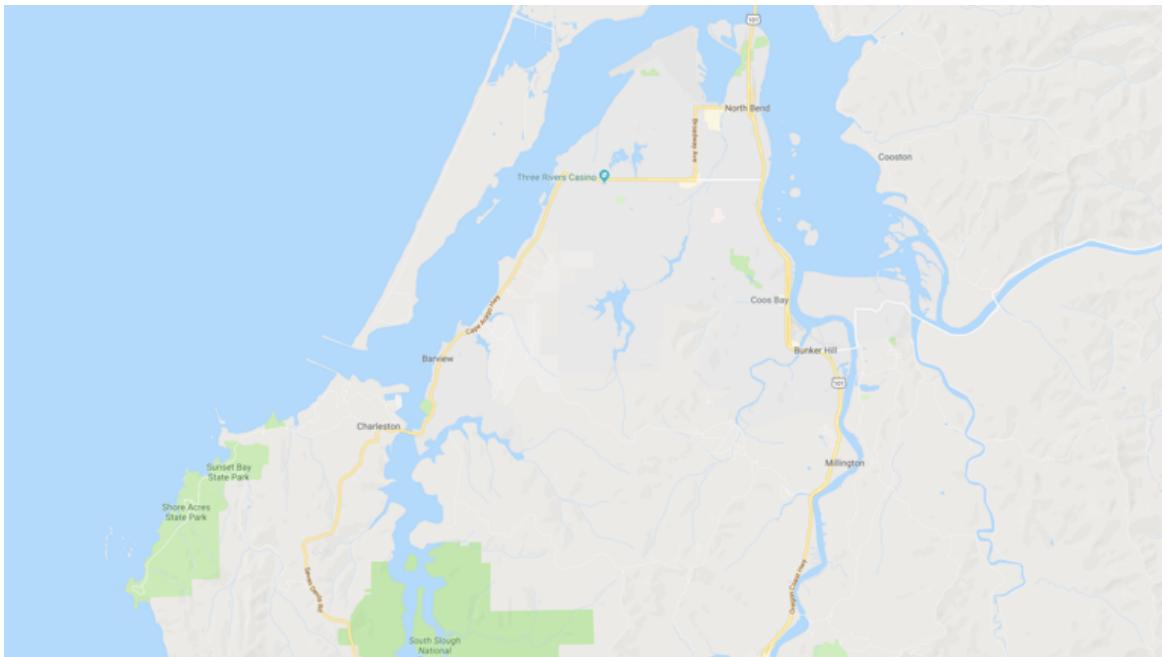
Zooming Into the Claim

Imagine Steve shows you this map and says, “Oregon’s coastline is too straight. I wish all coastlines were less straight so that they could all have a bay!”



Resist the temptation to argue back, “You’re wrong, bays are stupid!” Hopefully, you’ve built up the habit of nailing down a claim’s specific meaning before trying to argue against it.

Steve is making a claim about “Oregon’s coastline”, which is a pretty abstract concept. In order to unpack the claim’s specific meaning, we have to zoom into the concept of a “coastline” and see it in more detail as this specific configuration of land and water:



From this perspective, a good first reply would be, "Well, Steve, what about Coos Bay over here? Are you happy with Oregon's coastline as long as Coos Bay is part of it, or do you still think it's too straight even though it has this bay?"

Notice that we can't predict how Steve will answer our specific clarifying question. So we never knew what Steve's words meant in the first place, did we? Now you can see why it wasn't yet productive for us to start arguing against him.

When you hear a claim that sounds meaningful, but isn't 100% concrete and specific, the first thing you want to do is zoom into its specifics. In many cases, you'll then find yourself disambiguating between multiple valid specific interpretations, like for Steve's claim that "Oregon's coastline is too straight".

In other cases, you'll discover that there was *no* specific meaning in the mind of the speaker, like in the case of Steve's claim that "Acme exploits its workers by paying them too little"—a staggering thing to discover.



TFW a statement unexpectedly turns out to have no specific meaning

“Startups should have more impact!”

Consider this excerpt from a recent series of [tweets](#) by Michael Seibel, CEO of the [Y Combinator](#) startup accelerator program:

Successful tech founders would have far better lives and legacies if they competed for happiness and impact instead of wealth and users/revenue.

We need to change [the] model from build a big company, get rich, and then starting a foundation...

To build a big company, get rich, and use the company's reach and power to make the world a better place.

When I first read these tweets, my impression was that Michael was providing useful suggestions that any founder could act on to make their startup more of a force for good. But then I activated my specificity powers...

Before elaborating on what I think is the failure of specificity on Michael’s part, I want to say that I really appreciate Michael and Y Combinator engaging with this topic in the first place. It would be easy for them to keep their head down and stick to their original wheelhouse of funding successful startups and making huge financial returns, but instead, YC repeatedly pushes the envelope into new areas such as founding [OpenAI](#) and creating their [Request for Carbon Removal Technologies](#). The Y Combinator community is an amazing group of smart and morally good people, and I’m proud to call myself a YC founder (my company [Relationship Hero](#) was in the YC Summer 2017 batch). Michael’s heart is in the right place to

suggest that startup founders may have certain underused mechanisms by which to make the world a better place.



That said... is there any coherent takeaway from this series of tweets, or not?

The key phrases seem to be that startup founders should **“compete for happiness and impact”** and **“use the company’s reach and power to make the world a better place”**.

It sounds meaningful, doesn’t it? But notice that it’s generically-worded and lacks any specific examples. This is a red flag.

Remember when you first heard Steve’s claim that “Acme exploits its workers by paying them too little”? At first, it sounded like a meaningful claim. But as we tried to nail down what it meant, it collapsed into nothing. Will the same thing happen here?

Specificity powers, activate! Form of: [Tweet reply](#)

What's a specific example, real or hypothetical, of a \$1B+ founder trading off less revenue for more impact?

Cuz at the \$1B+ level, competing for impact may look indistinguishable from competing for revenue.

E.g. Elon Musk companies have huge impact and huge valuations.

Let’s consider a specific example of a startup founder who is highly successful: Elon Musk and his company SpaceX, currently valued at \$33B. The company’s mission statement is proudly displayed at the top of their [about](#) page:

SpaceX designs, manufactures and launches advanced rockets and spacecraft. The company was founded in 2002 to revolutionize space technology, with the ultimate goal of enabling people to live on other planets.

What I love about SpaceX is that everything they do follows from Elon Musk’s original goal of making human life multiplanetary. Check out this incredible [post](#) by Tim Urban to understand Elon’s plan in detail. Elon’s 20-year playbook is breathtaking:

1. **Identify a major problem in the world**

A single catastrophic event on Earth can permanently wipe out the human species

2. **Propose a method of fixing it**

Colonize other planets, starting with Mars

3. **Design a self-sustaining company or organization to get it done**

Invent reusable rockets to drop the price per launch, then dominate the \$27B/yr market for space launches

I would enthusiastically advise any founder to follow Elon’s playbook, as long as they have the stomach to commit to it for 20+ years.



So how does this relate to Michael's tweets? I believe my advice to "follow Elon's playbook" constitutes a specific example of Michael's suggestion to "use the company's reach and power to make the world a better place".

But here's the thing: Elon's playbook is something you have to do before you found the company. First you have to identify a major problem in the world, then you come up with a plan to start a certain type of company. How do you apply Michael's advice once you've already got a company?

To see what I mean, let's pick another specific example of a successful founder: Drew Houston and Dropbox (\$11B market cap). We know that Michael wants Drew to "compete for happiness and impact" and to "use the company's reach and power to make the world a better place". But what does that mean here? What specific advice would Michael have for Drew?



Let's brainstorm some possible ideas for specific actions that Michael might want Drew to take:

- Change Dropbox's mission to something that has more impact on happiness
- Donate 10% of Dropbox's profits to efforts to reduce world hunger
- Give all Dropbox employees two months of paid vacation each year

I know, these are just stabs in the dark, because we need to talk about specifics somehow. Did Michael really mean any of these? The ones about charity and employee benefits seem too obvious. Let's explore the possibility that Michael might be recommending that Dropbox change its mission.

Here's Dropbox's current mission from their [about](#) page:

We're here to unleash the world's creative energy by designing a more enlightened way of working.

Seems like a nice mission that helps the world, right? I use Dropbox myself and can confirm that the product makes my life a little better. So would Michael say that Dropbox is an example of "competing for happiness and impact"?

If so, then it would have been really helpful if Michael had written in one of his tweets, "I mean like how Dropbox is unleashing the world's creative energy". Mentioning Dropbox, or

any other specific example, would have really clarified what Michael is talking about.

And if Dropbox's current mission *isn't* what Michael is calling for, then how would Dropbox need to change it in order to better "compete for happiness and impact"? For instance, would it help if they tack on "and we guarantee that anyone can have access to cloud storage regardless of their ability to pay for it", or not?

Notice how this parallels my conversation with Steve about Acme. We begin with what sounds like a meaningful exhortation: *Companies should compete for happiness and impact instead of wealth and users/revenue! Acme shouldn't exploit its workers!* But when we reach for specifics, we suddenly find ourselves grasping at straws. I showed three specific guesses of what Michael's advice *could* mean for Drew, but we have no idea what it *does* mean, if anything.

Imagine that the CEO of Acme wanted to take Steve's advice about how not to exploit workers. He'd be in the same situation as Drew from Dropbox: confused about the specifics of what his company was supposedly doing wrong, to begin with.

Once you've mastered the power of specificity, you'll see this kind of thing everywhere: a statement that at first sounds full of substance, but then turns out to actually be empty. And the clearest warning sign is the absence of specific examples.

Next post: [How Specificity Works](#)

Utility ≠ Reward

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This essay is an adaptation of a talk I gave at the [Human-Aligned AI Summer School 2019](#) about [our work on mesa-optimisation](#). My goal here is to write an informal, accessible and intuitive introduction to the worry that we describe in our full-length report.

I will skip most of the detailed analysis from our report, and encourage the curious reader to follow up this essay with [our sequence](#) or [report](#).

The essay has six parts:

Two distinctions draws the foundational distinctions between “optimised” and “optimising”, and between utility and reward.

What objectives? discusses the behavioral and internal approaches to understanding objectives of ML systems.

Why worry? outlines the risk posed by the utility ≠ reward gap.

Mesa-optimisers introduces our language for analysing this worry.

An alignment agenda sketches different alignment problems presented by these ideas, and suggests transparency and interpretability as a way to solve them.

Where does this leave us? summarises the essay and suggests where to look next.

The views expressed here are my own, and do not necessarily reflect those of my coauthors or MIRI. While I wrote this essay in first person, all of the core ideas are the fruit of an equal collaboration between Joar Skalse, Chris van Merwijk, Evan Hubinger and myself. I wish to thank Chris and Joar for long discussions and input as I was writing my talk, and all three, as well as Jaime Sevilla Molina, for thoughtful comments on this essay.

≈3300 words.

Two distinctions

I wish to draw a distinction which I think is crucial for clarity about AI alignment, yet is rarely drawn. That distinction is between the reward signal of a reinforcement learning (RL) agent and its “utility function”^[1]. That is to say, it is not in general true that the policy of an RL agent is optimising for its reward. To explain what I mean by this, I will first draw another distinction, between “optimised” and “optimising”. These distinctions lie at the core of our mesa-optimisation framework.

It’s helpful to begin with an analogy. Viewed abstractly, biological evolution is an optimisation process that searches through configurations of matter to find ones that are good at replication. Humans are a product of this optimisation process, and so we

are to some extent good at replicating. Yet we don't care, by and large, about replication in itself.

Many things we care about *look* like replication. One might be motivated by starting a family, or by having a legacy, or by similar closely related things. But those are not replication itself. If we cared about replication directly, gamete donation would be a far more mainstream practice than it is, for instance.

Thus I want to distinguish the objective of the selection pressure that produced humans from the objectives that humans pursue. Humans were selected for replication, so we are good replicators. This includes having goals that correlate with replication. But it is plain that we are not motivated by replication itself. As a slogan, though we are optimised for replication, we aren't optimising for replication.

Another clear case where "optimised" and "optimising" come apart are "dumb" artifacts like [bottle caps](#). They can be heavily optimised for some purpose without optimising for anything at all.

These examples support the first distinction I want to make: **optimised ≠ optimising**. They also illustrate how this distinction is important in two ways:

1. A system optimised for an objective need not be pursuing any objectives itself.
(As illustrated by bottle caps.)
2. The objective a system pursues isn't determined by the objective it was optimised for. (As illustrated by humans.)

The reason I draw this distinction is to ask the following question:

Our machine learning models are optimised for some loss or reward. But what are they optimising for, if anything? Are they like bottle caps, or like humans, or neither?

| | | <i>optimised for</i> | <i>optimising for</i> |
|------------|-------|----------------------|-----------------------------------|
| Bottle cap | | keeping water in | n/a |
| Human | | replication | wealth, pleasure, fulfilment, ... |
| RL agent | | reward | ??? |

In other words, do RL agents have goals? And if so, what are they?

These questions are hard, and I don't think we have good answers to any of them. In any case, it would be premature, in light of the optimised ≠ optimising distinction, to conclude that a trained RL agent is optimising for its reward signal.

Certainly, the RL agent (understood as the agent's policy representation, since that is the part that does all of the interesting decision-making) is optimised for performance on its reward function. But in the same way that humans are optimised for replication, but are optimising for our own goals, a policy that was selected for its performance on reward may in fact have its own internally-represented goals, only indirectly linked to the intended reward. A pithy way to put this point is to say that **utility ≠ reward**, if we want to call the objective a system is optimising its "utility". (This is by way of

metaphor – I don't suggest that we must model RL agents as expected utility maximizers.)

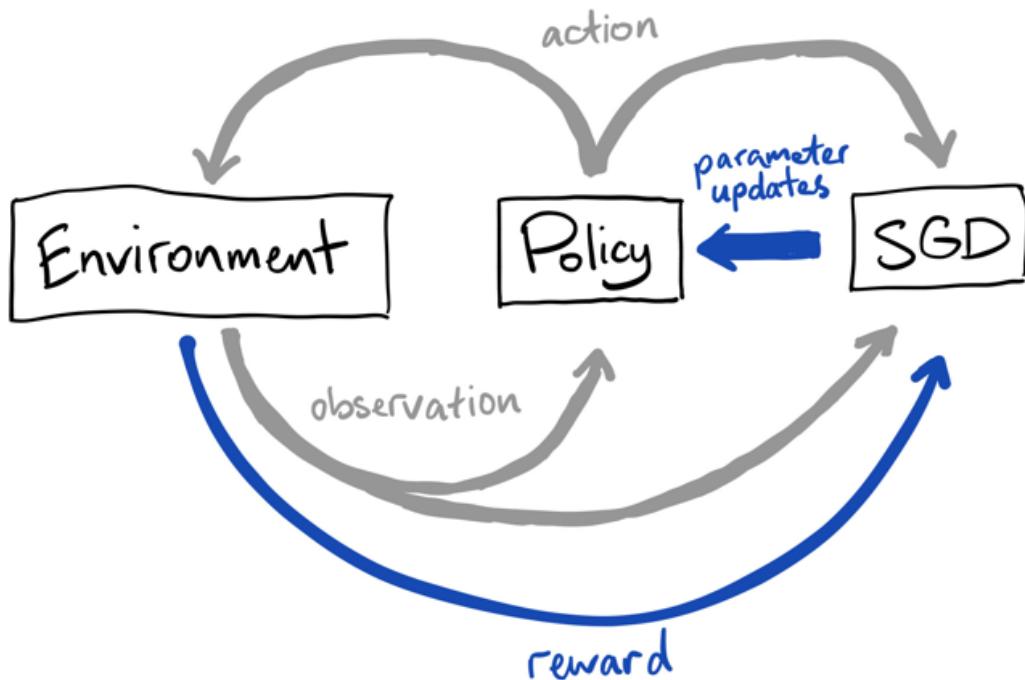
Let's make this more concrete with an example. Say that we train an RL agent to perform well on a set of mazes. Reward is given for finding and reaching the exit door in each maze (which happens to always be red). Then we freeze its policy and transfer the agent to a new environment set for testing. In the new mazes, the exit doors are blue, and red distractor objects are scattered elsewhere in the maze. What might the agent do in the new environment?

Three things might happen.

1. It might generalise: the agent could solve the new mazes just as well, reaching the exit and ignoring the distractors.
2. It might break under the distributional shift: the agent, unused to the blue doors and weirdly-shaped distractor objects, could start twitching or walking into walls, and thus fails to reach the exit.
3. But it might also fail to generalise in a more interesting way: the agent could fail to reach the exit, but could instead robustly and competently find the red distractor in each maze we put it in.

To the extent that it's meaningful to talk about the agent's goals, the contrast between the first and third cases suggests that those goals depend only on its policy, and are distinct from its reward signal. It is tempting to say that the objective of the first agent is reaching doors; that the objective of the third agent is to reach red things. It does not matter that in both cases, the policy was optimised to reach doors.

This makes sense if we consider how information about the reward gets into the policy:



For any given action, the policy's decision is made independently of the reward signal. The reward is only used (standardly, at least) to optimise the policy between actions. So the reward function can't be the policy's objective – one cannot be pursuing something one has no direct access to. At best, we can hope that whatever objective the learned policy has access to is an accurate representation of the reward. But the two can come apart, so we must draw a distinction between the reward itself and the policy's internal objective representation.

To recap: whether an AI system is goal-directed or not is not trivially answered by the fact that it was constructed to optimise an objective. To say that is to fail to draw the optimised \neq optimising distinction. If we then take seriously goal-directedness in AI systems, then we must draw a distinction between the AI's internal learned objective and the objective it was trained on; that is, draw the utility \neq reward distinction.

What objectives?

I've been talking about the objective of the RL agent, or its "utility", as if it is an intuitively sensible object. But what actually is it, and how can we know it? In a given training setup, we know the reward. How do we figure out the utility?

Intuitively, the idea of the internal goal being pursued by a learned system feels compelling to me. Yet right now, we don't have any good ways to make the intuition precise – figuring out how to do that is an important open question. As we start thinking about how to make progress, there are at least two approaches we can take: what I'd call the behavioural approach and the internal approach.

Taking the behavioural approach, we look at how decisions made by a system systematically lead to certain outcomes. We then infer objectives from studying those decisions and outcomes, treating the system as a black box. For example, we could apply Inverse Reinforcement Learning to our trained agents. Eliezer's [formalisation of optimisation power](#) also seems to follow this approach.

Or, we can peer inside the system, trying to understand the algorithm implemented by it. This is the internal approach. The goal is to achieve a mechanistic model that is abstract enough to be useful, but still grounded in the agent's inner workings. Interpretability and transparency research take this approach generally, though as far as I can tell, the specific question of objectives has not yet seen much attention.

It's unclear whether one approach is better, as both potentially offer useful tools. At present, I am more enthusiastic about the internal approach, both philosophically and as a research direction. Philosophically, I am more excited about it because understanding a model's decision-making feels more explanatory^[2] than making generalisations about its behaviour. As a research direction, it has potential for empirically-grounded insights which might scale to future prosaic AI systems. Additionally, there is the possibility of low-hanging fruit, as this space appears underexplored.

Why worry?

Utility and reward are distinct. So what? If a system is truly optimised for an objective, determining its internal motivation is an unimportant academic debate. Only its real-world performance matters, not the correct interpretation of its internals. And if the performance is optimal, then isn't our work done?

In practice, we don't get to optimise performance completely. We want to generalise from limited training data, and we want our systems to be robust to situations not foreseen in training. This means that we don't get to have a model that's perfectly optimised for the thing we actually want. We don't get optimality on the full deployment distribution complete with unexpected situations. At best, we know that the system is optimal on the training distribution. In this case, knowing whether the internal objective of the system matches the objective we selected it for becomes crucial, as if the system's [capabilities generalise while its internal goal is misaligned](#), bad things can happen.

Say that we prove, somehow, that optimising the world with respect to some objective is safe and useful, and that we can train an RL agent using that objective as reward. The utility ≠ reward distinction means that even in that ideal scenario, we are still not done with alignment. We still need to figure out a way to actually install that objective (and not a different objective that still results in optimal performance in training) into our agent. Otherwise, we risk creating an AI that appears to work correctly in training, but which is revealed to be pursuing a different goal when an unusual situation happens in deployment. So long as we don't understand how objectives work inside agents, and how we can influence those objectives, we cannot be certain of the safety of any system we build, even if we literally somehow have a proof that the reward it was trained on was "correct".

Will highly-capable AIs be goal-directed? I don't know for sure, and it seems hard to gather evidence about this, but my guess is yes. Detailed discussion is beyond our scope, but I invite the interested reader to look at some arguments about this that we present in [section 2](#) of the report. I also endorse Rohin Shah's [Will Humans Build Goal-Directed Agents?](#).

All this opens the possibility for misalignment between reward and utility. Are there reasons to believe the two will actually come apart? By default, I expect them to. Ambiguity and underdetermination of reward mean that there are many distinct objectives that all result in the same behaviour in training, but which can disagree in testing. Think of the maze agent, whose reward in training could mean "go to red things" or "go to doors", or a combination of the two. For reasons of bounded rationality, I also expect pressures for learning proxies for the reward instead of the true reward, when such proxies are available. Think of humans, whose goals are largely proxies for reproductive success, rather than replication itself. (This was a very brief overview; [section 3](#) of our report examines this question in depth, and expands on these points more.)

The second reason these ideas matter is that we might not want goal-directedness at all. Maybe we just want tool AI, or AI services, or some other kind of non-agentic AI. Then, we want to be certain that our AI is not somehow goal-directed in a way that would cause trouble off-distribution. This could happen without us building it in – after all, evolution didn't set out to make goal-directed systems. Goal-directedness just turned out to be a good feature to include in its replicators. Likewise, it may be that goal-directedness is a performance-boosting feature in classifiers, so powerful optimisation techniques would create goal-directed classifiers. Yet perhaps we are willing to take the performance hit in exchange for ensuring our AI is non-agentic. Right now, we don't even get to choose, because we don't know when systems are goal-directed, nor how to influence learning processes to avoid learning goal-directedness.

Taking a step back, there is something fundamentally concerning about all this.

We don't understand our AIs' objectives, and we don't know how to set them.

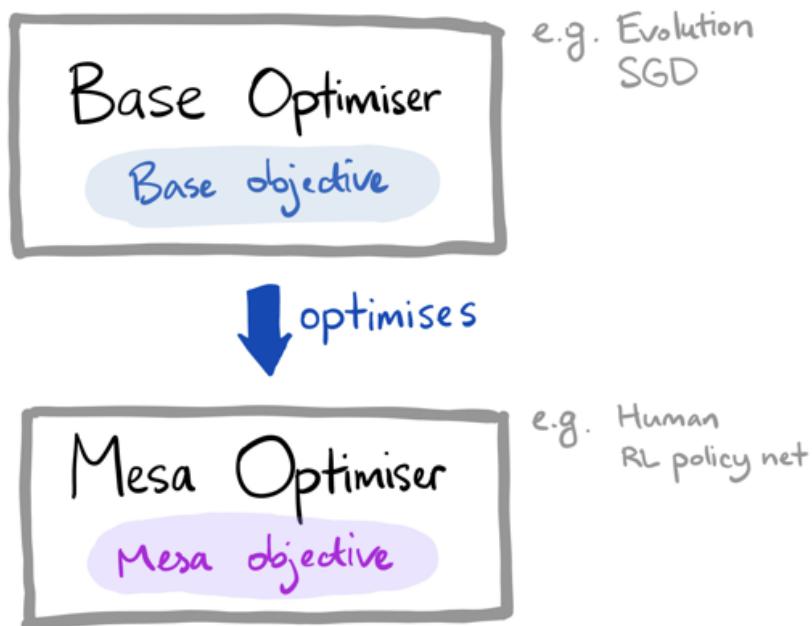
I don't think this phrase should ring true in a world where we hope to build friendly AI. Yet today, to my ears, it does. I think that is a good reason to look more into this question, whether to solve it or to assure ourselves that the situation is less bad than it sounds.

Mesa-optimisers

This worry is the subject of our report. The framework of mesa-optimisation is a language for talking about goal-directed systems under the influence of optimisation processes, and about the objectives involved.

A part of me is worried that the terminology invites viewing mesa-optimisers as a description of a very specific failure mode, instead of as a language for the general worry described above. I don't know to what degree this misconception occurs in practice, but I wish to preempt it here anyway. (I want data on this, so please leave a comment if you had confusions about this after reading the original report.)

In brief, our terms describe the relationship between a system doing some optimisation (the base optimiser, e.g.: evolution, SGD), and a goal-directed system (the mesa-optimiser, e.g.: human, ML model) that is being optimised by that first system. The objective of the base optimiser is the base objective; the internal objective of the mesa-optimiser is the mesa-objective.



("Mesa" as a Greek word that means something like the opposite of "meta". The reason we use "mesa" is to highlight that the mesa-optimiser is an optimiser that is itself being optimised by another optimiser. It is a kind of dual to a meta-optimiser, which is an optimiser that is itself optimising another optimiser. There is [contention](#) around whether this is good Greek, however.

While we're on the topic of terms, "inner optimiser" is a confusing term that we used in the past in the same way as "mesa-optimiser". It did not accurately reflect the

concept, and has been retired in favour of the current terminology. Please use "mesa-optimiser" instead.)

I see the word "optimiser" in "mesa-optimiser" as a way of capturing goal-directedness, rather than a commitment to some kind of (utility-)maximising structure. What feels important to me in a mesa-optimiser is its goal-directedness, not the fact it is an optimiser. A goal-directed system which isn't taking strictly optimal actions (but which is still competent at pursuing its mesa-objective) is still worrying.

Optimisation could be a good way to model goal-directedness—though I don't think you gain that much, conceptually, from that model—but equally, it seems plausible that some other approach we have not yet explored could work better. So I myself read the "optimiser" in "mesa-optimiser" analogously to how I accept treating humans as optimisers; as a metaphor, more than anything else.

I am not sure that mesa-optimisation is the best possible framing of these concerns. I would welcome more work that attempts to untangle these ideas, and to improve our concepts.

An alignment agenda

There are at least three alignment-related ideas prompted by this worry.

The first is **unintended optimisation**. How do we ensure that systems that are not supposed to be goal-directed actually end up being not-goal-directed?

The second is to factor alignment into **inner alignment** and **outer alignment**. If we expect our AIs to be goal-directed, we can view alignment as a two-step process. First, ensure outer alignment between humans and the base objective of the AI training setup, and then ensure inner alignment between the base objective and the mesa-objective of the resulting system. The former involves finding low-impact, corrigible, aligned with human preferences, or otherwise desirable reward functions, and has been the focus of much of the progress made by the alignment community so far. The latter involves figuring out learned goals, interpretability, and a whole host of other potential approaches that have not yet seen much popularity in alignment research.

The third is something I want to call **end-to-end alignment**. It's not obvious that alignment must factor in the way described above. There is room for trying to set up training in such a way to guarantee a friendly mesa-objective somehow without matching it to a friendly base-objective. That is: to align the AI directly to its human operator, instead of aligning the AI to the reward, and the reward to the human. It's unclear how this kind of approach would work in practice, but this is something I would like to see explored more. I am drawn to staying focused on what we actually care about (the mesa-objective) and treating other features as merely levers that influence the outcome.

We must make progress on at least one of these problems if we want to guarantee the safety of prosaic AI. If we don't want goal-directed AI, we need to reliably prevent unintended optimisation. Otherwise, we want to solve either inner and outer alignment, or end-to-end alignment. Success at any of these requires a better understanding of goal-directedness in ML systems, and a better idea of how to control the emergence and nature of learned objectives.

More broadly, it seems that taking these worries seriously will require us to develop better tools for looking inside our AI systems and understanding how they work. In light of these concerns I feel pessimistic about relying solely on black-box alignment techniques. I want to be able to reason about what sort of algorithm is actually implemented by a powerful learned system if I am to feel comfortable deploying it.

Right now, learned systems are (with maybe the exception of feature representation in vision) more-or-less hopelessly opaque to us. Not just in terms of goals, which is the topic here—most aspects of their cognition and decision-making are obscure. The alignment concern about objectives that I am presenting here is just one argument for why we should take this obscurity seriously; there may be other risks hiding in our poor understanding of AI inner workings.

Where does this leave us?

In summary, whether a learned system is pursuing any objective is far from a trivial question. It is also not trivially true that a system optimised for achieving high reward is optimising for reward.

This means that with our current techniques and understanding, we don't get to know or control what objective a learned system is pursuing. This matters because in unusual situations, it is that objective that will determine the system's behaviour. If that objective mismatches the base objective, bad things can happen. More broadly, our ignorance about the cognition of current systems does not bode well for our prospects at understanding cognition in more capable systems.

This forms a substantial hole in our prospects at aligning prosaic AI. What sort of work would help patch this hole? Here are some candidates:

- Empirical work. Distilling examples of goal-directed systems and creating convincing scaled-down examples of inner alignment failures, like the maze agent example.
- Philosophical, deconfusion and theoretical work. Improving our conceptual frameworks about goal-directedness. This is a promising place for philosophers to make technical contributions.
- Interpretability and transparency. Getting better tools for understanding decision-making, cognition and goal-representation in ML systems.

These feel to me like the most direct attacks on the problem. I also think there could be relevant work to be done in verification, adversarial training, and even psychology and neuroscience (I have in mind something like a review of how these processes are understood in humans and animals, though that might come up with nothing useful), and likely in many more areas: this list is not intended to be exhaustive.

While the present state of our understanding feels inadequate, I can see promising research directions. This leaves me hopeful that we can make substantial progress, however confusing these questions appear today.

1. By “utility”, I mean something like “the goal pursued by a system”, in the way that it’s used in decision theory. In this post, I am using this word loosely, so I don’t give a precise definition. In general, however, clarity on what exactly “utility” means for an RL agent is an important open question. ↪

2. Perhaps the intuition I have is a distant cousin to [the distinction drawn by Einstein between principle and constructive theories](#). The internal approach seems more like a “constructive theory” of objectives. [←](#)

The Inadequacy of Current Science (Novum Organum Book 1: 1-37)

This is the third post in the [Novum Organum sequence](#). For context, see the [sequence introduction](#).

We have used Francis Bacon's *Novum Organum* in the version presented at www.earlymoderntexts.com. Translated by and copyright to [Jonathan Bennett](#). Prepared for LessWrong by [Ruby](#).

Ruby's Reading Guide

Novum Organum is organized as two books each containing numbered "aphorisms." These vary in length from three lines to sixteen pages. Titles of posts in this sequence, e.g. *Idols of the Mind Pt. 1*, are my own and do not appear in the original.

While the translator, Bennett, encloses his editorial remarks in a single pair of [brackets], I have enclosed mine in a [[double pair of brackets]].

Bennett's Reading Guide

[Brackets] enclose editorial explanations. Small ·dots· enclose material that has been added, but can be read as though it were part of the original text. Occasional •bullets, and also indenting of passages that are not quotations, are meant as aids to grasping the structure of a sentence or a thought. Every four-point ellipsis indicates the omission of a brief passage that seems to present more difficulty than it is worth. Longer omissions are reported between brackets in normal-sized type.

Aphorism Concerning the Interpretation of Nature: Book 1: 1-37

by Francis Bacon

- 1.** Man, being nature's servant and interpreter, is limited in what he can do and understand by what he has observed of the course of nature—directly observing it or inferring things ·from what he has observed·. Beyond that he doesn't know anything and can't do anything.
- 2.** Not much can be achieved by the naked hand or by the unaided intellect. Tasks are carried through by tools and helps, and the intellect needs them as much as the hand does. And just as the hand's tools either •give motion or •guide it, so ·in a comparable way· the mind's tools either •point the intellect in the direction it should go or • offer warnings.
- 3.** Human knowledge and human power meet at a point; for where the cause isn't known the effect can't be produced. The only way to command nature is to obey it; and something that functions as the •cause in thinking about a process functions as the •rule in the process itself.
- 4.** All that man can do to bring something about is to put natural bodies together or to pull them away from one another. The rest is done by nature working within.
- 5.** The mechanic, the mathematician, the physician, the alchemist and the magician have all rubbed up against nature in their activities; but so far they haven't tried hard and haven't achieved much.
- 6.** If something has never yet been done, it would be absurd and self-contradictory to expect to achieve it other than through means that have never yet been tried.

[[Similar: [not every change is an improvement, but every improvement is a change.](#)]]
- 7.** If we go by the contents of •books and by •manufactured products, the mind and the hand seem to have had an enormous number of offspring. But all that variety consists in very fine-grained special cases of, and derivatives from, a few things that were already known; *not* in a large number of fundamental propositions.
- 8.** Moreover, the works that have already been achieved owe more to chance and experiment than to disciplined sciences; for the sciences we have now are merely pretty arrangements of things already discovered, not ways of making discoveries or pointers to new achievements.
- 9.** Nearly all the things that go wrong in the sciences have a single cause and root, namely: while wrongly admiring and praising the powers of the human mind, we don't look for true helps for it.
- 10.** Nature is much subtler than are our senses and intellect; so that all those elegant meditations, theorizing and defensive moves that men indulge in are crazy—except that no-one pays attention to them.

[Bacon often uses a word meaning 'subtle' in the sense of 'fine-grained, delicately complex'; no one current English word will serve.]

[[An especially good example of this point, that nature is far more "subtle" than our senses and mind, is the generally counterintuitive fact that our universe runs on [quantum mechanics](#). In [Can You Prove Two Particles Are Identical?](#), Eliezer points to this weird aspects of reality that one is very unlikely to discover without the empiricism/appropriate tools and methodology which Bacon is advocating for.]]

11. just as the sciences that we now have are useless for devising new inventions, the logic that we now have is useless for discovering new sciences.

[Bacon here uses *inventio* in two of its senses, as = 'invent' and as = 'discover'.]

12. The logic now in use serves to •fix and stabilize errors based on the ideas of the vulgar, rather than to •search for truth. So it does more harm than good.

[[The next few aphorisms dealing with syllogism and axioms are made with reference to the [Aristotelian 'scientific method.'](#) In that classical approach, a few real-world examples are used to derive high-level [universal rules or laws](#) which are then operated on with logic to derive further conclusions. See [this comment below](#) for more detail.]]

The leap from a few examples to high-level general principles is what Bacon is calling out when in **19** he speaks of 'swooping up' from particulars to general axioms. This is in contrast to his gradual, incremental, *inductive* method that starts with limited statements of rule and only slowly generalizes as more data is accumulated.]]

13. The syllogism isn't brought to bear on the •basic principles of the sciences; it *is* applied to •intermediate axioms, but nothing comes of this because the syllogism is no match for nature's subtlety. It constrains what you can assent to, but not what can happen.

[[These remarks bear resemblance to those in [The Parable of Hemlock](#).]]

14. A •syllogism consists of •propositions, which consist of •words, which are stand-ins [*tesserae*, literally = 'tickets'] for •notions. So the root of the trouble is this: If the notions are confused, having been sloppily abstracted from the facts, nothing that is built on them can be firm. So our only hope lies in true induction.

15. There is no soundness in •our• notions, whether in logic or in natural science. These are not sound notions:

- substance, quality, acting, undergoing, being;

And these are even less sound:

- heavy, light, dense, rare, moist, dry, generation, corruption, attraction, repulsion, element, matter, form

and so on; all of those are fantastical and ill-defined.

['Rare' = 'opposite of dense'. Generation is the coming into existence of living things; corruption is rotting or falling to pieces, and so refers to the going out of existence of

living things. For the next sentence: a ‘lowest species’ is one that doesn’t further divide into subspecies.]

16. ·Our· notions of the lowest species (*man, dog, dove*) and of the immediate perceptions of the senses (*hot, cold, black, white*) don’t seriously mislead us; yet even they are sometimes confusing because of how matter flows and things interact. As for all the other notions that men have adopted—they are mere aberrations, not being caused by things through the right kind of abstraction.

17. The way •axioms are constructed is as wilful and wayward as the abstractions through which •notions are formed. I say this even about the principles that result from vulgar induction, but much more about the axioms and less basic propositions that the syllogism spawns.

18. The discoveries that have been made in the sciences up to now lie close to vulgar notions, scarcely beneath the surface. If we are to penetrate into nature’s inner and further recesses, we’ll need •a safer and surer method for deriving notions as well as axioms from things, as well as •an altogether better and more certain way of conducting intellectual operations.

19. There are and *can be* only two ways of searching into and discovering truth. **(1)** One of them starts with the senses and particular events and swoops straight up from them to the most general axioms; on the basis of these, taken as unshakably true principles, it proceeds to judgment and to the discovery of intermediate axioms. This is the way that people follow now. **(2)** The other derives axioms from the senses and particular events in a gradual and unbroken ascent, ·going through the intermediate axioms and· arriving *finally* at the most general axioms. This is the true way, but no-one has tried it.

[[Reminder that ‘dialectics’ is generally Bacon’s term for logic, but he is seemingly specifically referring to the [logic and processes followed in Aristotle’s methods](#).]]

20. When the intellect is left to itself it takes the same way—namely **(1)**—that it does when following the rules of dialectics. For the mind loves to leap up to generalities and come to rest with them; so it doesn’t take long for it to become sick of experiment. But this evil, ·though it is present both in natural science and in dialectics·, is worse in dialectics because of the ordered solemnity of its disputationes.

21. When the intellect of a sober, patient, and grave mind is left to itself (especially in a mind that isn’t held back by accepted doctrines), it ventures a little way along **(2)** the right path; but it doesn’t get far, because without guidance and help it isn’t up to the task, and is quite unfit to overcome the obscurity of things.

22. Both ways set out from the senses and particular events, and come to rest in the most general propositions; yet they are enormously different. For one of them **(1)** merely glances in passing at experiments and particular events, whereas the other **(2)** stays among them and examines them with proper respect. One **(1)** proceeds immediately to laying down certain abstract and useless generalities, whereas the other **(2)** rises by step by step to what is truly better known by nature.

[In calling something ‘known to nature’ Bacon means that it is a general law of nature; ‘better known by nature’ could mean ‘a more general law of nature’ or ‘a generality that is more completely lawlike’.]

23. There is a great difference between •the *idols* of the human mind and •the *ideas* of God's mind—that is, between •certain empty beliefs and •the true seals [= 'signs of authenticity'] and marks that we have found in created things.

24. There's no way that axioms •established by argumentation could help us in the discovery of new things, because the subtlety of nature is many times greater than the subtlety of argument. But axioms •abstracted from particulars in the proper way often herald the discovery of new particulars and point them out, thereby returning the sciences to their active status.

25. The axioms that are now in use are mostly made so that they *just* cover the items from which they arise, namely thin and common-or-garden experiences and a few particulars of the commonest sorts, so it is no wonder if they don't lead to new particulars. ·And it's not only the axioms, but also the way they are handled, that is defective·. If some unexpected counter-example happens to turn up, the axiom is rescued and preserved by some frivolous distinction, rather than (the truer course) being amended.

[[Once upon a time, the philosophers of Plato's Academy claimed that the best definition of human was a "featherless biped". Diogenes of Sinope, also called Diogenes the Cynic, is said to have promptly exhibited a plucked chicken and declared "Here is Plato's man." The Platonists promptly changed their definition to "a featherless biped with broad nails". - [Similarity Clusters](#)]]

26. To help me get my ideas across, I have generally used different labels for human reason's two ways of approaching nature: the customary way I describe as *anticipating nature* (because it is rash and premature) [Note from the preface: throughout this work, 'anticipation' means something like 'second-guessing, getting ahead of the data, jumping the gun'. Bacon means it to sound rash and risky; no one current English word does the job.] and the way that draws conclusions from facts in the right way I describe as *interpreting* nature.

27. Anticipations are a firm enough basis for consent, for even if men all went mad in the same way they might agree one with another well enough.

[[consent = agreement]]

28. Indeed, anticipations have much more power to win assent than interpretations do. They are inferred from a few instances, mostly of familiar kinds, so that they immediately brush past the intellect and fill the imagination; whereas interpretations are gathered from very various and widely dispersed facts, so that they can't suddenly strike the intellect, and must seem weird and hard to swallow—rather like the mysteries of faith.

[[Bacon appears to be saying that the easy, quick, rash science is easy to convince people of since it has low [inferential distance](#) owing to it being derived from a few familiar examples; in contrast, his difficult and true science built on many observations and facts has high inferential distance, causing it to seem strange and weird.]]

29. Anticipations and dialectics have their place in sciences based on opinions and dogmas, because in those sciences the aim is to be master of •what people believe but not of •the facts.

30. Even if all the brains of all the ages come together, collaborate and share their results, no great progress will ever be made in science by means of anticipations. That is because errors that are rooted in the first moves that the mind makes can't be cured later on by remedial action, however brilliant.

31. It is pointless to expect any great advances in science from grafting new things onto old. If we don't want to go around in circles for ever, making 'progress' that is so small as be almost negligible, we must make a fresh start with deep foundations.

[‘Fresh start’ translates *instauratio*, from the verb *instauro* = ‘make a fresh start (on a ceremony that has been wrongly performed)’. Bacon planned a six-part work on science and its philosophy and methods, which he called his *Instauratio magna*—his Great Fresh Start. There are other informal mentions of fresh starts in **38** and **129**, and the Great Fresh Start is referred to in **92** and each of **115–117**. Bacon died six years after publishing the present work. It is Part 2 of the Great Fresh Start, and the only Part he completed.]

32. This is not to attack the honour of the ancient authors or indeed of anyone else, because I am comparing not •intelligences or •competences but •ways ·of proceeding in the sciences·; and the role I have taken on is that of a guide, not a judge.

33. This must be said outright: anticipations (the kind of reasoning that is now in use) can't pass judgment on my method or on discoveries arising from it; for I can't be called on to submit to the sentence of a tribunal which is itself on trial!

34. It won't be easy for me to deliver and explain my message, for things that are in themselves *new* will be understood on analogy with things that are *old*.

35. Borgia said that when the French marched into Italy they came with chalk in their hands to •mark out their lodgings, not with weapons to •force their way in. Similarly, I want my doctrine to enter quietly into the minds that are fit to receive it and have room for it. ·Forcing my way in with weapons, so to speak, won't work· because refutations—and more generally *arguments pro and con*—can't be employed when what's at stake is a difference of view about first principles, notions, and even forms of demonstration.

36. There remains for me only one way of getting my message across. It is a simple way, namely this: I must lead you to the particular events themselves, and to the order in which they occur; and you for your part must force yourself for a while to lay aside your •notions and start to familiarize yourself with •facts.

37. Those who deny that anything can be known for sure •start off their thinking in something like my way, but where they •end up is utterly different from and opposed to where I end up. They say that *nothing can be known*, period. I say that *not much can be known about nature by the method that is now in use*. And then they go on to destroy the authority of the senses and the intellect, whereas I devise and supply helps for them.

The next post in the sequence, Book 1: 38-52 (Idols of the Mind Pt. 1), will be posted Tuesday, September 24 at latest by 4:00pm PDT.

Novum Organum: Introduction



In light of its value as a rationalist text, its historical influence on the progress of science, and its general expression of the philosophy and vision which guides LessWrong 2.0, the moderation team has seen fit to publish Novum Organum as a LessWrong sequence. (Image: the engraved title page.)

Quotes in this post are from Francis Bacon's Novum Organum in the version by Jonathan Bennett presented at www.earlymoderntexts.com

In 1620, Francis Bacon's [Novum Organum](#) was published. Though the work might be succinctly described as Bacon's views on empiricism and [inductivism](#), it is far more than a list of experimental steps to be followed. It is an entire epistemology and philosophy—possibly *the* epistemology and philosophy which underlay the [Scientific Revolution](#).

Bacon was damning of the science of his time and preceding centuries. He saw the [pseudo-empirical syllogistic paradigm](#) as deeply flawed and incapable of making progress.

If those doctrines ·of the ancient Greeks· hadn't been so utterly like a plant torn up by its roots, and had remained attached to and nourished by the womb of nature, the state of affairs that we have seen to obtain for two thousand years—namely *the sciences stayed in the place where they began, hardly changing, not getting any additions worth mentioning, thriving best in the hands of their first founders and declining from then on*—would never have come about. (74) [1]

He also believed that the unaided human mind was incapable of getting far on its own.

Nearly all the things that go wrong in the sciences have a single cause and root, namely: while wrongly admiring and praising the powers of the human mind, we don't look for true helps for it. (9)

Not much can be achieved by the naked hand or by the unaided intellect. Tasks are carried through by tools and helps, and the intellect needs them as much as the hand does. (2)

When the intellect of a sober, patient, and grave mind is left to itself (especially in a mind that isn't held back by accepted doctrines), it ventures a little way along the right path; but it doesn't get far, because without guidance and help it isn't up to the task, and is quite unfit to overcome the obscurity of things. (21)

Nonetheless, he was optimistic that if the old doctrines were abandoned, *idols of the mind* (i.e., biases, fallacies, and confusions) were cleared out, and his precise, careful empirical method was followed by a community of scholars, then no knowledge was out of reach and humanity would eventually achieve all of the most splendid discoveries.

Until now men haven't lingered long with •experience; they have brushed past it on their way to the ingenious •theorizings on which they have wasted unthinkable amounts of time. But if we had someone at hand who could answer our questions of the form 'What are the facts about this matter?', it wouldn't take many years for us to discover all causes and complete every science. (112)

The human mind is fallible and flawed—"like a [distorting mirror](#)," Bacon says—yet its biases can be overcome. Through adherence to properly looking at the world, such that if "[the road from the senses to the intellect](#) [is] well defended with walls along each side," then a scientific community can figure out the world and even reach [Utopia](#).

This a decidedly LessWrong worldview.

Indeed, by my reading, Bacon possessed in some form a large number of concepts employed on LessWrong, not limited to: [confirmation bias](#), [motivated cognition](#), [the bottom line](#), [mind-projection fallacy](#), [positive bias](#), [entangled evidence](#), [carving reality at its joints](#), [fake causality](#), [worshipping ignorance](#), [idea inoculation](#), [the surprisingly detailedness of reality](#), [inferential distance](#), [incentives](#), and [dissolving confused language](#). He even spoke of the [appropriate degrees of certainty](#) for each stage of an inquiry and deliberately used [epistemic statuses](#)!

Novum Organum was Bacon's monumental attempt to explain all of the above: how and why the existing scientific methods were entirely broken, why nobody had noticed until then, what the alternative paradigm was, and a vision for a community of scholars and institutions which could help discover all scientific truths.

Covering biases and empiricism as it does, Novum Organum is highly instructive as a rationalist text. Yet why read Bacon when we've got [the Sequences](#), [Codex](#), and the

rest of modern LessWrong? I answer that it's worthwhile because there's a focus and immediacy to a text whose author wasn't writing abstractly, but direly wanted to redirect all the scientific efforts of his time to be more productive.

There's an impressiveness to someone grappling with how to do science at a point when so much less was known about the world. Compared to us, Bacon's time was one of extreme mystery. Recall that he was writing before Boyle, Newton, Maxwell, or Darwin. He did not have access to theories of thermodynamics, electromagnetism, evolution, or atomic physics. They hadn't even invented the mercury thermometer in his time. He earnestly tried to figure out simply "what is heat?" and by use of his meticulous empiricism correctly inferred it was just something to do with motion—150 years before phlogiston theory was laid to rest and with access to only primitive air-based thermometers!

We get to look back and point to all that modern science has done over the centuries to make us feel enthusiastic. Four hundred years ago, Bacon's enthusiasm came entirely from his ability to look forward.

There is also perhaps a validation of the LessWrong worldview to be found in Bacon. Bacon was a symbolic figure of the Scientific Revolution. Inspirational to the Royal Society and many others. Historical credit allocation is hard, but it seems more likely than not that Bacon gets a good deal of credit in bringing about the Scientific Revolution. Seemingly, many of the same ideas that we cherish now were read by the scholars who first read Bacon and kicked off the modern scientific era. If only people hadn't stopped reading Bacon in the original after a few generations.

Beyond his instruction in biases and empiricism, Bacon is an inspiration to the LessWrong 2.0 project [2] for his visions of how infrastructure and community are key to intellectual progress. Bacon saw intellectual progress as a [technological](#) [3] and [collaborative](#) endeavor, exactly as LessWrong 2.0 does.

At the technologies for individual thinking level, Bacon writes:

Not much can be achieved by the naked hand or by the unaided intellect. Tasks are carried through by tools and helps, and the intellect needs them as much as the hand does. And just as the hand's tools either give motion or guide it, so ·in a comparable way· the mind's tools either point the intellect in the direction it should go or offer warnings. (2)

Bacon is further adamant that the process of science requires people to write their work down and share it. Perhaps this is obvious now, but Bacon was writing before the first scientific journal, indeed, he is credited as a major inspiration for the Royal Society whose [philosophical transactions were the first scientific journal](#).

Even after we have acquired and have ready at hand a store of natural history and experimental results such as is required for the work of the intellect, or of philosophy, still that is not enough. The intellect is far from being able to retain all this material in memory and recall it at will, any more than a man could keep a diary all in his head. Yet until now there has been more thinking than writing about discovery procedures—experimentation hasn't yet become literate! But a discovery isn't worth much if it isn't ·planned and reported· in writing; and when this becomes the standard practice, better things can be hoped for from experimental procedures that have at last been made literate. (101)

Yet another point, maybe, obvious to us now: the work of science can be split up among people.

Unlike the work of sheerly thinking up hypotheses, proper scientific work can be done collaboratively; the best way is for men's efforts (especially in collecting experimental results) to be exerted separately and then brought together. Men will begin to know their strength only when they go this way—with one taking charge of one thing and another of another, instead of all doing all the same things.

(113)

Though Bacon's greatest reference to collaborating and institution for knowledge perhaps comes from his utopian novel, [New Atlantis](#). One character describes the fictional institution of [Solomon's House](#):

Ye shall understand (my dear friends) that amongst the excellent acts of that king, one above all hath the pre-eminence. It was the erection and institution of an Order or Society, which we call *Salomon's House*; the noblest foundation (as we think) that ever was upon the earth; and the lanthorn of this kingdom. It is dedicated to the study of the works and creatures of God. Some think it beareth the founder's name a little corrupted, as if it should be Solamona's House. But the records write it as it is spoken. So as I take it to be denominated of the king of the Hebrews, which is famous with you, and no stranger to us.

The novel goes into great depth about how the institution functions and all the roles different individuals play in the scientific process. According to Wikipedia, it is this vision which inspired Samuel Hartlib and Robert Boyle to found the Royal Society.

To conclude this introduction, I'll mention that Novum Organum is actually part two of six from Bacon's much larger, never-completed work, [Instauratio Magna](#). The title is usually translated as *The Great Instauration* yet [Bennett](#) (whose translation of Novum Organum we are posting) translates it as *The Great Fresh Start*. Seems fitting to Bacon's intentions.

It is pointless to expect any great advances in science from grafting new things onto old. If we don't want to go around in circles for ever, making 'progress' that is so small as be almost negligible, we must make a fresh start with deep foundations. (31)

Given the Scientific Revolution got going in earnest around his lifetime, I dare say he got what we he asked for.

[1] Novum Organum consists two books each containing "aphorisms" which range in length from three lines to sixteen pages. A bold number on its own refers to an aphorism from Book 1 by default or Book 2 where the context is very clear. When unclear, aphorisms are referenced by a leading 1- or 2- to disambiguate, e.g 2-13 is the 13th aphorism in Book 2.

[2] Usually, we now call ourselves simply "LessWrong" but it feels important to disambiguate here since I cannot make claims to the vision for original LessWrong as founded in 2009 by Eliezer. It does seem clear that Eliezer was not influenced by Bacon in the same way that Habryka (LessWrong 2.0's team lead and core founder) has been.

[3] By *technological* I refer broadly to the creation of knowledge and tools that can be used for a specific purpose, including things like methodologies and procedures, not just physical artifacts. I would call a set of techniques for debiasing one's thinking and likewise training for how to moderate an online forum as both examples of technologies.

Age gaps and Birth order: Reanalysis

This post follows on from my [previous post](#) detailing some areas where I was unable to reproduce Scott's [analysis](#) of how the age gap between siblings modifies the SSC Birth order effect. I suggest you read that post first but here's the summary:

I attempted to reproduce Scott's analysis of [Birth order effect vs Age gap](#). I found that:

There appeared to be an error in graphs 2 & 3 where people with one sibling were counted when they shouldn't have been (graph 2) or were counted twice (graph 3)

Comparing oldest children to youngest children causes a bias in the results which can be prevented by comparing oldest children to 2nd oldest children

I was unable to reproduce Scott's result on people reporting 0 year age gap - I get a non-significant 58% older siblings compared to Scott's 70%. I was unable to discover the cause of the difference.

Summary

I reanalysed how sibling age gap modifies the SSC birth order effect. I found that:

The birth order effect is relatively steady for the first 4-8 years of age gap at about 70% respondents being the firstborn vs secondborn. For larger age gaps the effect reduces. There is insufficient evidence to conclude how long this reduction takes or whether the effect is completely removed at very large age gaps.

2 other trends were noted in the data but evidence for them was not strong:

- The reduction may not be the same (or might disappear) for larger families
- Birth order effect may be lower at 1 year age gap vs 2-7 year age gap

Considering competing theories on the cause of the Birth order effect, two theories fit the data well:

- Intra-family dynamics
- Decreased parental investment

And three theories fit the data poorly:

- Changed parental strategies
- Maternal antibodies
- Maternal vitamin deficiencies

Introduction

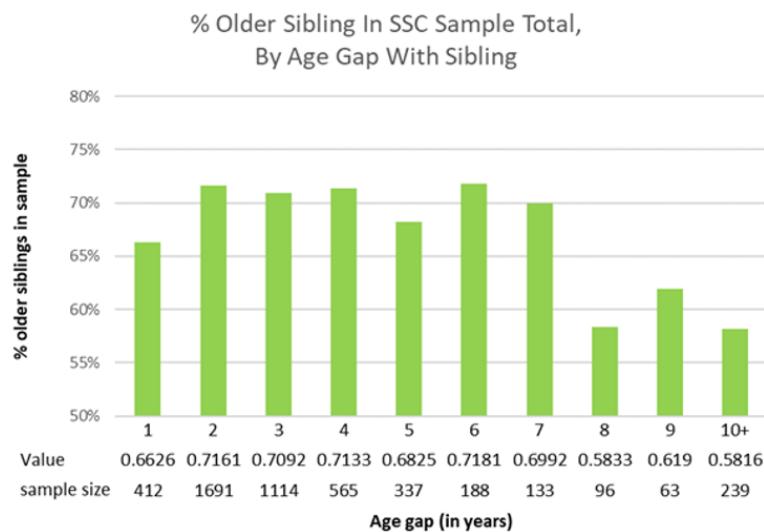
The original reason for me looking at this data was to analyse whether the data support a sudden drop between years 7 and 8 or whether there is an alternative explanation which fits the data.

I will note here that I'm not a trained statistician and am using this as practice of Bayesian model comparison, inspired by johnswentworth's recent [model comparison](#) sequence. I'd say I'm 80% confident in my broad conclusions, less so in the specifics - I'd be fairly confident there are a couple of errors lurking in here somewhere.

Analysis: All family sizes combined

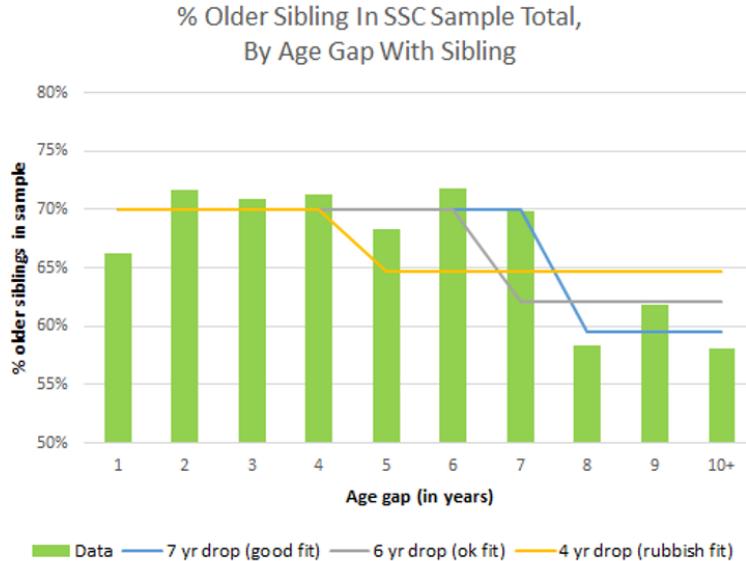
Is there a sudden drop after 7 years?

Getting back to the data, here's the result that I'm going to focus on, comparing 1st to 2nd children in all family sizes:



Eyeballing the graph makes the sudden drop after 7 years look like the most natural explanation. However, we had no reason, a priori, to think that a 7 year age gap would have any special significance - a drop could have happened after 1 or 10 years for all we knew.

If we model a sudden drop after 6 or 8 years the model starts to match the data significantly less well, any further away from 7 than that and the model performs really poorly. Although a general "sudden drop" model has a high maximum likelihood at 7 years, the overall model likelihood is lower due to the lower likelihoods for other drop years.

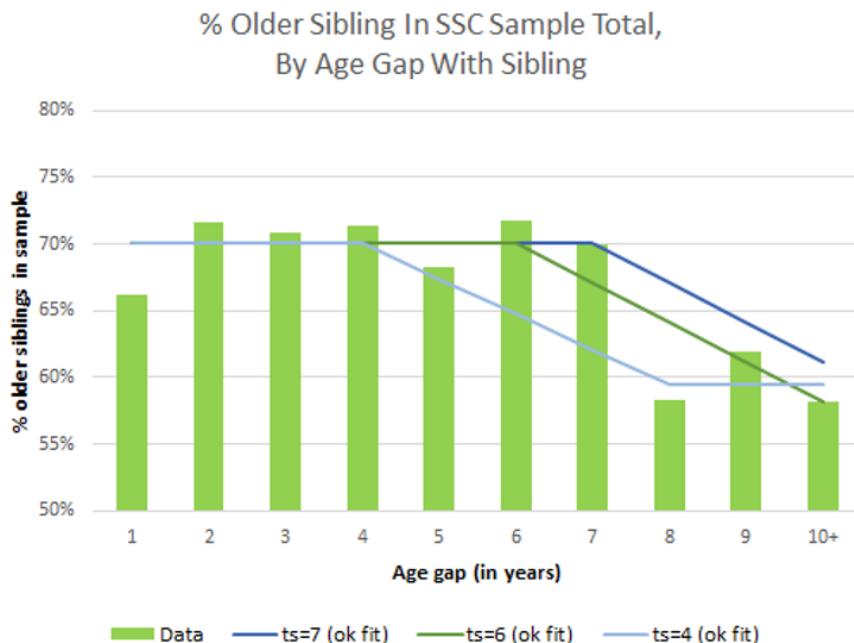


General slope model

Imagine a model which is similar to a sudden drop model but the drop is ramped down over a number of years. The model is defined by 4 parameters - percentage oldest sibling before the ramp (p_0), percentage oldest sibling after the ramp (p_1), at what age gap the ramp starts (t_s) and over how many years the ramp occurs (t_r).

The sudden drop model is nested within this model - where $t_r = 0$.

A gentler slope doesn't match the data as closely as a sudden drop but is less harshly penalised over a range of ramp start locations. The graph below shows what some $t_r = 4$ years ramps might look like.



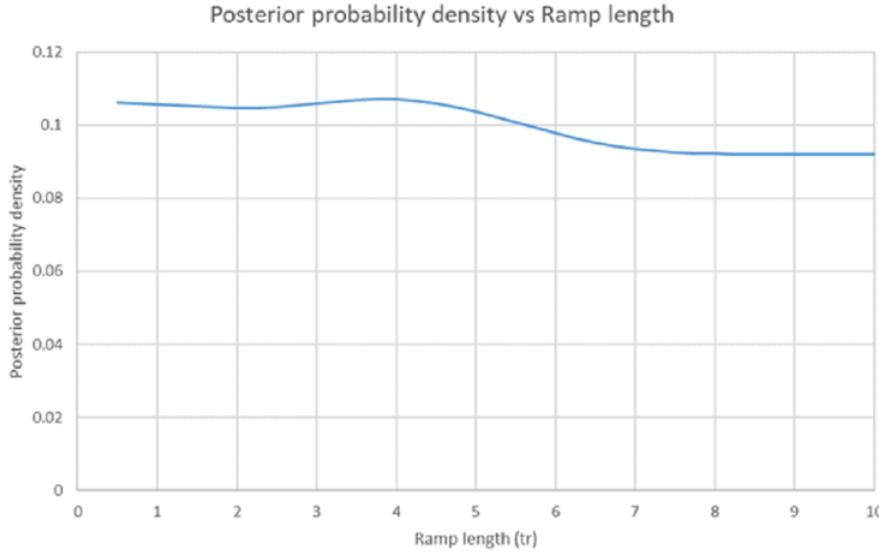
Ramp timing and length

To find out which ramp lengths fit the data best I integrate (numerically) across the first 3 parameters in this model (p_0, p_1, t_s) to find which value of the 4th parameter (t_r) predicts the data the best – how sudden is the drop?

(Notes:

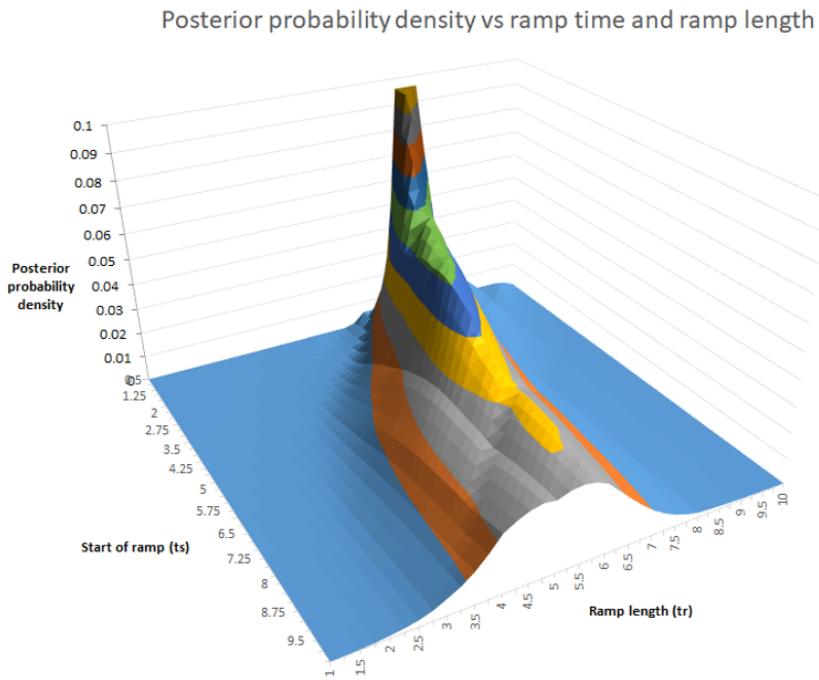
For this analysis I haven't grouped the 10+ year age gaps together but used the actual values for the age gaps.

For all calculations in this post I assume a uniform prior across a reasonable range for each parameter.)



Surprisingly, the likelihood is fairly flat over a large range of slope lengths – everything between 0 and 10 years is within a Bayes factor of 1.15 of each other.

To see what's happening, let's integrate over the first two parameters (p_0 and p_1) and plot likelihood against ramp length (t_r) and start (t_s).

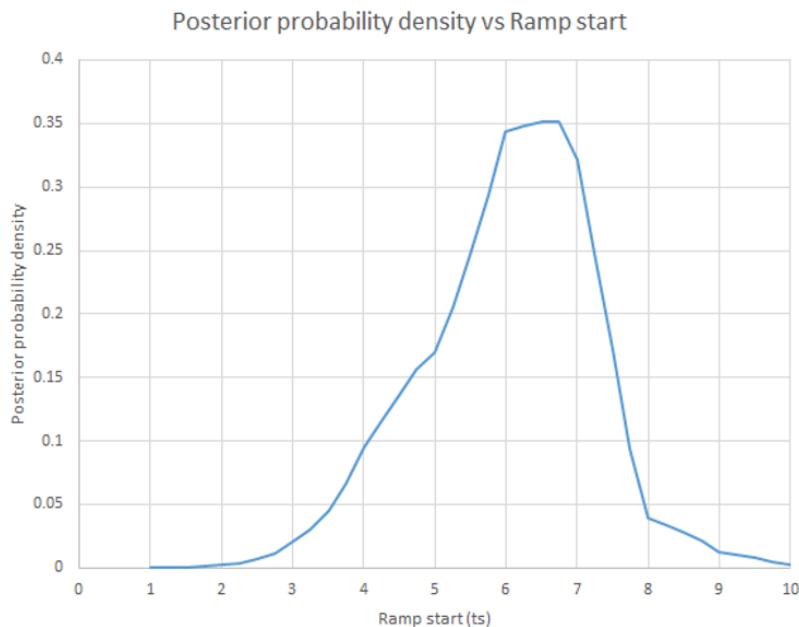


This shows a maximum value at $t_r = 0$, $t_s = 7$ – the sudden drop after 7 years which is so visually noticeable in the data.

However, if you follow the line along $t_r = 0$ (back of the graph), there is only a small range of t_s values which have a high likelihood. Looking instead along $t_r = 5$, the maximum likelihood is lower (~33% lower), but there is a larger range of t_s values which provide a fairly high likelihood. The decrease in maximum likelihood is almost exactly cancelled out by the increase in the width of the distribution.

So a sudden drop predicts the data approximately as well as a more gradual drop.

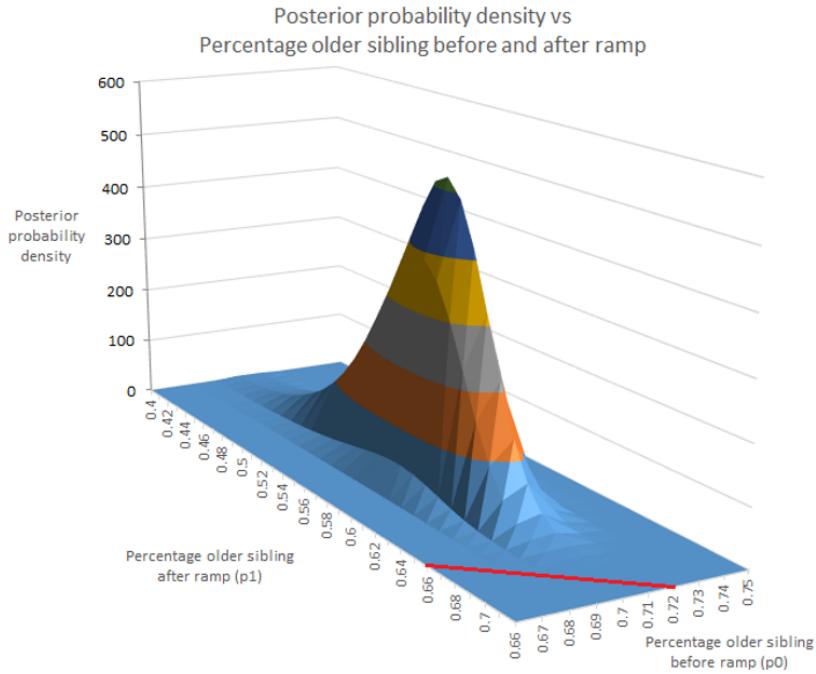
We can also integrate across t_r to find the posterior probability of the various t_s values.



I'm going to describe this as the ramp starting between 4 and 8 years.

Percentage oldest children before and after drop

I also integrated over t_r and t_m in order to see how likelihood varied with p_0 and p_1 .



p_0 is very precisely defined between 0.70 and 0.71.

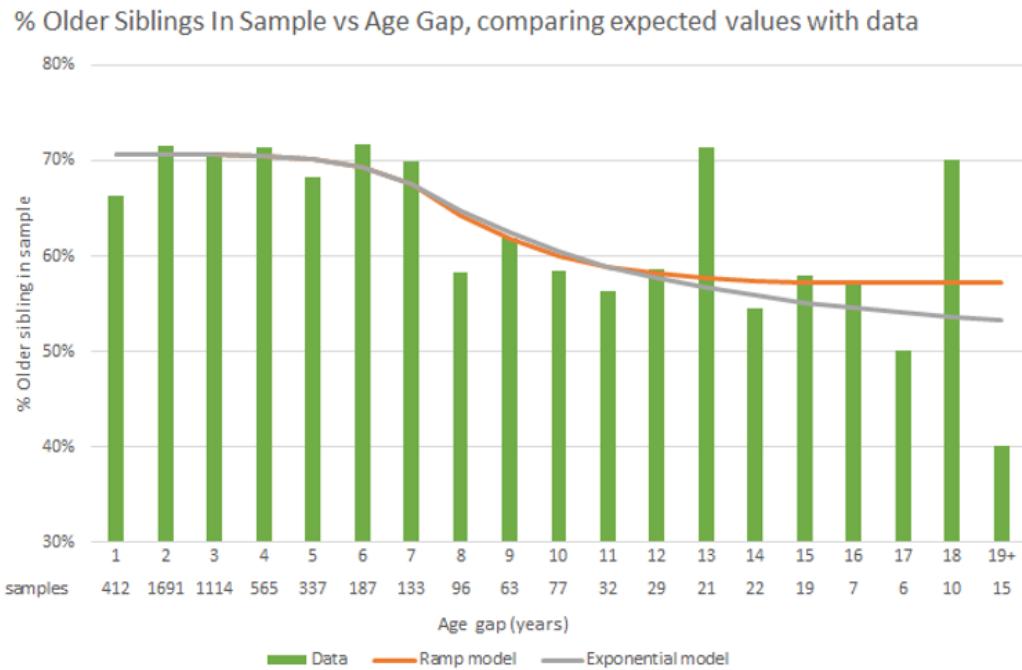
p_1 can take a large variety of values, between ~0.49 & 0.62 (90% CI).

In reality, the Birth order effect might decrease relatively fast to start with and then more slowly as oldest and second oldest children approach parity. This is probably the kind of thing which we would expect in real life but which can't be recreated with the ramp model.

I created an exponential decay model (with a delay in the decay starting) to test whether this might be the case and it got a slightly higher overall likelihood than the general ramp model (Bayes factor 1.5). The start of the decline was in the region 3-8 years, similar to the ramp model. The maximum likelihood half-life was 5 years although this could be anywhere between 1.2-11 years (90% CI).

Expected values

Using these models I calculated expected values for Birth order effect vs age gap.



This looks fairly sensible to me. There is a gradual start to the slope, becoming steeper into about year 8 and then shallowing out as we get closer to parity between older and younger siblings.

At larger age gaps the two models diverge which is due to a combination of the differing priors implied by the models and the sparsity of data points in this region - the likelihood isn't sufficient to overcome the prior.

Comparison to constant birth effect model

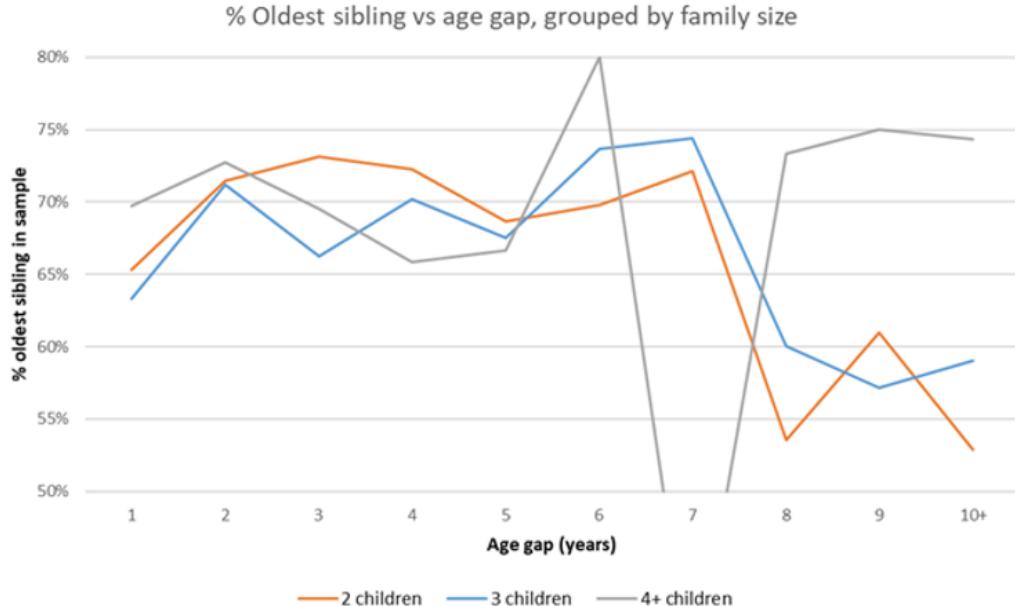
I also compared the general ramp model to a constant Birth order effect model. The ramp model was preferred over the constant model by a Bayes factor of ~1,000.

A constant model is actually nested within the ramp model where $p_0 = p_1$ (and t_r, t_m become meaningless). This is illustrated by the red line on the likelihood vs p_0 & p_1 graph where the low likelihood can be seen.

Analysis: Different sized families

I mentioned in my previous post that it appeared that the drop was present in sibships of 2 but not in sibships of 3+.

Breaking this down further, we can compare this effect for sibships of 2, sibships of 3 and sibships of 4+ (any further breakdown causes the sample sizes to get too small).



(The very low value at 7 year age gap for 4+ children is only a sample size of 11 so don't take it too seriously!)

Here it appears that the drop-off in birth effect for large age gaps between first and second children happens in sibships of 2 or 3 but doesn't happen in sibships of 4+.

Although the number of samples in the 4+ group with >7 year age gap is only 64, the difference between 2-3 and 4+ sibships is significant at $p<0.05$ (two-tailed t-test).

This seems an odd phenomenon. Would having extra siblings cause the birth order effect between the oldest 2 siblings to remain high for large age gaps?

Seeing something weird like this in my data causes me to ask "how many things might I have spotted during my work on this project, if they had coincidentally shown a weird looking result?" - when adjusting for post-hoc multiple hypothesis testing I should adjust not just for the tests that I did but also for the tests I didn't do just because nothing looked odd.

In this case the answer is quite a lot so $p<0.05$ is probably not strict enough and my best bet would be that this data occurred by coincidence.

That's all a bit hand-wavey so I tried to calculate the Bayes factor comparing:

A general ramp model for all family sizes

vs

A general ramp model for families of 2 & 3 children combined with a shallower (or no) ramp for families of 4+ children (Only p_1 was changed between the family sizes)

The latter was preferred by a factor of 5. If I were to include other numbers of children when the change might have happened or possibility that the change happens

gradually as family size got bigger then this factor would change but that would start getting way too complicated for me!

I still don't really believe this an actual effect but if someone has an explanation of what might cause this then I'm all ears.

Possible lower birth order effect for 1 year age gap

One other thing which I noticed is the lower Birth order effect for age gaps of 1 year as compared to gaps of 2-7 years (0.66 vs 0.71 oldest siblings). A quick calculation suggests Bayes factor comes out at 2 in favour of the Birth order effect being lower at 1 year age gap compared it being constant across 1-7 year age gaps.

Note in this case that although the Bayes factor isn't huge, it seems like this is the kind of thing which might actually happen (some of the potential causes would give this a decent prior - see section below for more discussion) so I'm much less inclined to just write this one off.

Comparing Explanations for Birth order effect

Scott [lists](#) 5 potential causes of the Birth order effect:

1. Intra-family competition
2. Decreased parental investment
3. Changed parenting strategies
4. Maternal antibodies
5. Maternal vitamin deficiencies

I've renamed 1 to "Intra-family dynamics" to include non-competitive interactions between siblings. A few people have mentioned other sibling dynamics which might cause a Birth order effect (e.g. [here](#)). The predictions of age gap effect from competitive vs non-competitive causes seem similar to me so I'll lump them together.

My thoughts for what each of the 5 potential causes would predict regarding age gap are given below. The conclusions for each potential cause end up being very similar to Scott's (after all that work!) except that there is no need to postulate anything especially significant about 7 years and that there may be a slight increase in birth order effect between 1 and 2 years age gap.

Intra-family dynamics

Prediction: Birth order effect remains roughly constant with small age gaps, with less effect as the gap gets larger.

Assessment: Findings match prediction well. 4-8 years seems reasonable for levels of interactions between siblings to start decreasing.

Potentially, for a small age gap, a very advanced younger sibling might act more like an older sibling meaning that the 1 year age gap birth effect would be lower. This feels slightly forced to me (I would think any such effect would be fairly small) but am curious what others think.

Decreased parental investment

Prediction: Birth order effect increases as age gap increases - the longer a firstborn is the only child the longer they benefit from 100% of their parents' attention. If the earliest years are the most important then birth order might not change after that critical period. Once older children are able to look after themselves, birth order effect might come down with larger age gaps.

Assessment: The increase in birth order effect between 1 and 2 years would match the theory, if parental investment is mostly important in the first two years. If older children start being able to look after themselves after 4-8 years then this would explain the drop in birth order effect after this time.

The match between the theory and result is good, although there are a couple of degrees of freedom to help match the prediction to the data. 4-8 years seems reasonable for children starting to look after themselves better but 2 years seems on the low side for a prediction of how long having extra attention is beneficial. Maybe between 2-5 years the two effects roughly cancel out?

Changed parenting strategies

Prediction: Age gap has minimal effect on Birth order effect.

Assessment: Prediction matches data poorly. It is possible that parental strategies start to reset towards firstborn strategies after longer age gaps but I wouldn't have put much of my probability mass on that option. There is a 5 year gap between my youngest children and I definitely didn't reset towards firstborn strategies, I suspect this would have still been true even for a much larger gap.

Maternal antibodies

Prediction: Age gap has minimal effect on Birth order effect. Generally you don't need top-ups of vaccines so presumably antibodies stick around indefinitely? Or is it your body's ability to make more? Anyway, Scott thinks this is unlikely and he's a doctor so I'll take his word for it.

Assessment: Prediction matches data poorly. My biology knowledge is too poor to know how likely a decrease in effectiveness after 4-8 years would be in this case.

Maternal vitamin deficiencies

Prediction: Very small age gaps have large effect. Birth order effect decreases rapidly for age gaps <3 years – my estimate for how long it might take to rebuild vitamin stockpiles.

Assessment: Prediction matches data poorly. 4-8 years seems way too long for vitamin stockpiles to *start* to build back up.

Conclusions

The SSC 2019 survey data support a constant, high, birth order effect (~2.4 oldest siblings for every 1 second oldest sibling) for age gaps <4-8 years. This is followed by a decline to a lower birth order effect at an undetermined rate. The decline does not necessarily completely remove any birth order effect although this may be the case for very large age gaps.

The data provide some evidence that:

- The reduction may not be the same (or might disappear) for larger families (4+ children)
- Birth order effect may be lower at 1 year age gap vs 2-7 year age gap

However the evidence for both of these points is relatively slim.

Intra-family dynamics and decreased parental investment predict the results well.

Changed parental strategies, maternal antibodies and maternal vitamin deficiencies do not predict the results well.

How Specificity Works

This is Part II of the [Specificity Sequence](#)

You saw what mayhem we brought forth when we activated the first power of specificity, the power to [demolish bad arguments](#), and hopefully your curiosity is piqued to see what'll happen when we activate all the other powers.

But first, let's pause here to ask: *How does specificity work?*

Consider this dialogue between Steve and one of his pals:

Steve: Information should be free!

Steve's Pal: Whoa, this is thought-provoking stuff. Ok, so, would you say you're advocating for digital socialism, or more like digital libertarianism?

Oh jeez. Not only is Steve's pal not pushing for Steve to be more specific, the pal is an enabler who pushes Steve to be *less* specific. They're climbing the ladder of abstraction the wrong way.

The Ladder of Abstraction

When you want to nail down a claim, the operative word is “down”: you want to bring the discussion down the [ladder of abstraction](#):



If Steve says, “Information should be free!” and I’m trying to understand what he means, here’s what I’d say:

Liron: Ok, why do you think I shouldn’t have had to pay Amazon for my paperback copy of *To Kill A Mockingbird*?

This way, I’m nosediving all the way down to the bottom rung of the ladder of abstraction. Down here, the conversation becomes grounded in the concrete language of everyday experience, with substantive statements like “Amazon charged my credit card \$6.99 and kicked back \$1.15 to Harper Lee’s estate.”

To Kill a Mockingbird, 50th Anniversary Edition The 50th Anniversary edition Edition

by Harper Lee (Author)

4.5 out of 5 stars 10,442 customer reviews

Kindle \$11.99 Audiobook from \$28.99 Hardcover \$17.67 Paperback \$3.68 - \$6.99 Other Sellers See all 255 versions

Buy used \$3.68

Buy new **\$6.99** ✓prime 23 New from \$6.99

FREE delivery: Wednesday, July 24 Details
Deliver to Liron - Saratoga Spr... 12866
In stock on July 20, 2019.
Order it now.
Ships from and sold by Amazon.com.
✓prime

Qty: 1 Add to Cart or 1-Click Checkout

Buy now with 1-Click®
Free shipping once available
This is a gift

More Buying Choices 52 used & new from \$6.99
23 New from \$6.99 | 29 Used from \$3.68 See All Buying Options

ISBN-13: 978-0099549482
ISBN-10: 9780099549482
Why is ISBN important? ▾

Have one to sell? Sell on Amazon

Add to List

Share Email Facebook Twitter Pinterest Embed

And how about that free shipping, Steve?

Having this kind of grounded discussion is usually more productive than having a flingfest of the higher level ballpit-words “information”, “freedom”, “socialism”, and “libertarianism”.

As Steve and I are hanging out on the bottom rung of the ladder of abstraction, talking through specific examples of information getting exchanged freely vs. non-freely, we'll be able to notice if certain features seem to remain constant across our chosen examples. For instance, we might discuss various hypothetical authors who yearn to write books for a living, and observe that all such authors can still have some plausible mechanism to earn money (running ads on their blogs?), even without directly charging for the privilege of reading their books.

After loading our brains with specific examples, we can then abstract over these examples, climb our way back up the ladder of abstraction, and put forth a generalized claim about whether “information should be free”. Since we've been careful to think about specific example scenarios that our claim applies to, we'll be putting forth a coherent and meaningful proposition - a claim that others can study under the magnifying lens of specificity, not an empty claim that gets demolished.

So when someone makes an abstract assertion like “information should be free”, the best thing you can do is hold their hand and guide them down the ladder of abstraction. I know it's tempting to skip that process and just attack or admire their original abstract claim. But show me two people discussing a topic in purely abstract terms, and I'll show you two people who are talking past each other.

How To Slide Downward

How do you take a concept and slide it down the ladder of abstraction to obtain a more specific concept? What mental operation must your brain perform?

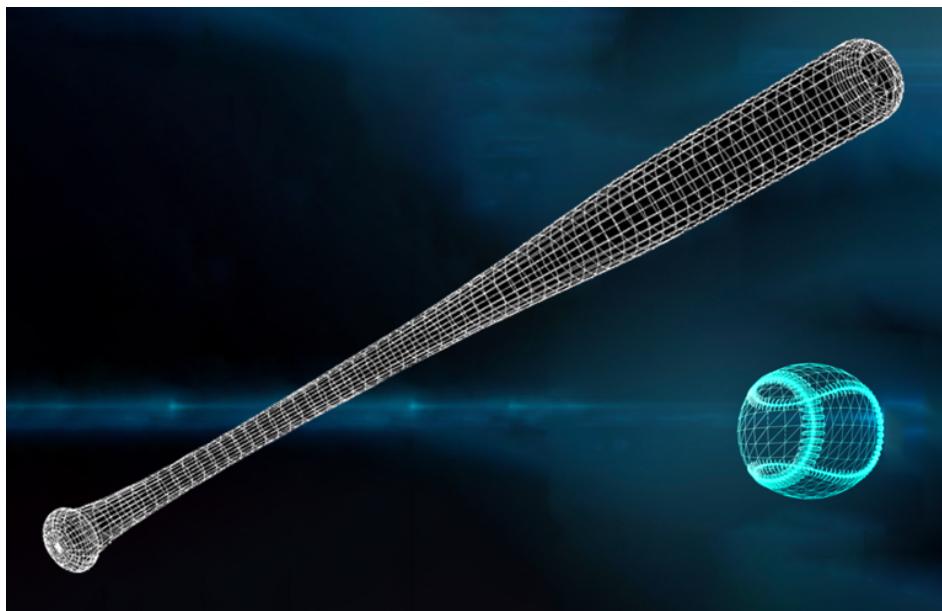
In [Replace the Symbol with the Substance](#), Eliezer explains how to do it with baseball terms:

You have to visualize. You have to make your mind's eye see the details, as though looking for the first time.

Is that a "bat"? No, it's a long, round, tapering, wooden rod, narrowing at one end so that a human can grasp and swing it.

Is that a "ball"? No, it's a leather-covered spheroid with a symmetrical stitching pattern, hard but not metal-hard, which someone can grasp and throw, or strike with the wooden rod, or catch.

Are those "bases"? No, they're fixed positions on a game field, that players try to run to as quickly as possible because of their safety within the game's artificial rules.



Or let's say we're discussing our country's education system. Eliezer would slide it down the ladder of abstraction like this:

Why are you going to "school"? To get an "education" ending in a "degree". Blank out the forbidden words and all their obvious synonyms, visualize the actual details, and you're much more likely to notice that "school" currently seems to consist of sitting next to bored teenagers listening to material you already know, that a "degree" is a piece of paper with some writing on it, and that "education" is forgetting the material as soon as you're tested on it.

Let's try it ourselves with the concept of a school "lecture". What do we get when we slide it down the ladder of abstraction?

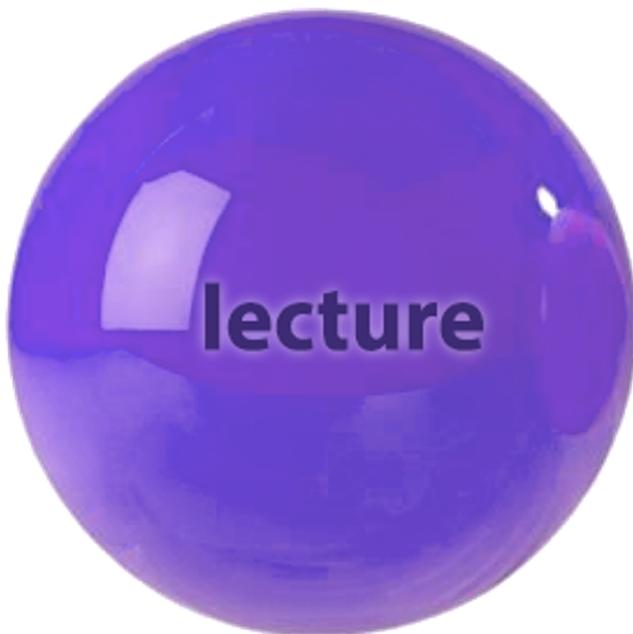
"A stage presentation of publicly-available educational material, hand-produced and performed by a professor who works at your educational institution, which you watch by locating yourself in a set building at a set meeting time, and which proceeds in a fixed order and at a fixed rate like broadcast television pre-YouTube."

Wow. When you hear it that way, it raises a lot of questions:

- How about using the best educational materials from the internet as a course's official materials? Those are surely better than what any professor in your school has ever performed.
- How about making it a standard expectation for students to consume lectures at their own pace, including taking advantage of the pause and speed up / slow down features?
- How about not forcing students to show up to a lecture hall at a specific time?

If people had never heard the word “lecture”, if people were always forced to talk about lectures via a specific description of what a lecture is... well, then they would have killed off lectures by now.

I believe the college lecture is only alive today because the word “lecture” is a protective abstraction-bubble.



I grabbed this out of Steve's ball pit

If you crack open the protective shell of “lecture” and press your nose close enough, you breathe in the stinky innards: “a stage presentation of publicly-available educational material”.

If everyone involved with the university system—administrators, professors, parents, students—were themselves cracking open the shell and taking a whiff of “lecture”, they would have noticed when the expiration date passed (the day YouTube went mainstream) and taken out the trash.

Instead, we've landed in a weird place where the concept of a “stage presentation of publicly-available educational material” is ridiculous on its face, while our ears still tell

us that the concept of a school “lecture” sounds pretty good.

Ground Your Terms

You probably know that to have a clear discussion (or just a clear thought), you need to define your terms. How do you define a term?

S. I. Hayakawa illustrates an attempt to define the term “red” by connecting it to concepts *higher up* the ladder of abstraction (h/t [Eliezer](#)):

“What is meant by the word red?”
“It’s a color.”
“What’s a color?”
“Why, it’s a quality things have.”
“What’s a quality?”

This approach of defining something by sliding it up the ladder of abstraction doesn’t feel productive. It might *help* define “red”, but it’s neither necessary nor sufficient to define “red”.

Similarly, when I asked Steve to define what “exploiting workers” means in regard to Acme, and he put forth “to use selfishly for one’s own ends”, we found ourselves no closer to understanding what the heck his point was supposed to be.

So how can we nail down “red”? How can we slide it *down* the ladder of abstraction? Hayakawa illustrates:

“What is meant by the word red?”
“Well, the next time you see some cars stopped at an intersection, look at the traffic light facing them. Also, you might go to the fire department and see how their trucks are painted.”



Now we’re getting somewhere. This is a good enough definition to satisfy someone who previously didn’t know what “red” means. Whenever we define a concept in this

manner, by sliding it down the ladder of abstraction, we can call it **grounding** the concept.

Grounding is easy enough to do. Just follow Eliezer's instructions from the previous section:

You have to visualize. You have to make your mind's eye see the details, as though looking for the first time.

For example: What is fire?

If you know chemistry, you might *define* it as "rapid oxidation accompanied by heat and, usually, light". But if you don't know chemistry, what would you say? Most people would give up.

Don't worry, just follow the instructions to ground it. Close your eyes and describe what you see:

"The bright orange heat and light that appears when I strike a match, and can sometimes be transferred to other things it touches, and keeps appearing as long as there is wood and air around it."

Concepts have many definitions, and not all of them are groundings. But in daily life, grounding a term is usually as good as precisely defining it, if not better. So when a term is confusing, just slide that sucker down the ladder of abstraction.

Effort and Risk Asymmetry

If you observe your own stream of consciousness in a discussion, you might feel it being gently buoyed up the ladder of abstraction. You might start a discussion with a few specific statements about firetrucks, but before you know it you're talking about redness and colors in general.

Why is that? If the most productive kind of discussion is grounded and concrete, then why do so many people seem to relish the experience of pontificating and arguing abstractly?

Eliezer says in [The 5-Second Level](#):

Over-abstraction happens because it's easy to be abstract. It's easier to say "red is a color" than to pause your thoughts for long enough to come up with the example of a stop sign. Abstraction is a path of least resistance, a form of mental laziness.

It seems our brain stores each concept with an easily-navigable pointer upward to its category (like red→color), but doesn't store easily-navigable pointers downward to specific examples (like red→firetruck). I'm not sure what larger aspect of the brain's architecture accounts for this, but here's one observation about why the two operations are asymmetrical:

When you slide a concept up, you remove information. When you slide it down, you add information, and that requires you to make an arbitrary choice with more degrees of freedom. I think there's a sense in which going up the ladder of abstraction is *safer*,

while going down is *riskier*: your underconstrained choice of specifics may leak information about you that your [elephant in the brain](#) is keen to monitor and filter.

Eliezer's description of how he intentionally sticks his neck out in arguments has always stayed with me ([source](#)), and I suspect it's related to why we find abstraction appealing:



I stick my neck out so that it can be chopped off if I'm wrong, and when I stick my neck out it stays stuck out, and if I have to withdraw it I'll do so as a visible concession. I may parry [...] but I at least endeavor not to dodge. Where I plant my standard, I have sent an invitation to capture that banner; and I'll stand by that invitation.

When you say something abstract, like "information should be free", the space of possible things you can mean is vast. You're not sticking your neck out, you're not affixing your neck to a precise location in claim-space, so you never have to fear that an opponent's sword might slash there.

Plus, as a bonus, vague statements make you sound smarter. You're signaling more intelligence and sophistication talking about "digital libertarianism" than about "buying a paperback on Amazon". That's why abstraction is appealing.

The upshot is that you'll have to make a sustained conscious effort to acquire the skill and habit of activating your specificity powers. But it'll be worth it.

Next post: [The Power to Judge Startup Ideas](#)

[Talk] Paul Christiano on his alignment taxonomy

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a linkpost for <https://www.youtube.com/watch?v=-vsYtevJ2bc>

Paul Christiano makes an amazing tree of subproblems in alignment. I saw this talk live and enjoyed it.

The Power to Be Emotionally Mature

This is Part VII of the [Specificity Sequence](#)

When [Steve](#) unknowingly lacks an understanding of what his claim that "Uber exploits its drivers" means, we can say that he's being intellectually immature. Similarly, when someone unknowingly lacks an understanding of the cause of an emotion they're expressing, we can say they're being *emotionally* immature.

Imagine a 7-year-old boy playing with his sister in the park at dusk, who doesn't realize that he's currently tired and cranky. Sis gives him a soft punch on the arm, and suddenly he bursts into tears, runs to his mom and complains that his sister was being mean to him.

Boy: Sis was mean to me!

Mom: What specifically did she do?

Boy: She punched me!

Mom: Was it a hard punch or just a mean punch?

Boy: It was just a mean punch!



This is the emotional equivalent of:

Steve: Uber exploits its drivers!

Liron: What specifically do they do?

Steve: They pay them too little!

Liron: Is Uber paying a typical driver less than that driver would earn if Uber didn't exist, or just too little to be non-exploitative?

Steve: Just too little to be non-exploitative!

If you want to be emotionally mature, there's a mental operation you can do which we'll call **grounding your emotions** because it's analogous to [grounding your terms](#).

Recall that grounding a term requires you to describe a specific stimulus that you'd mentally classify within that term, like when we grounded the concept of "red" by saying it's something you call a traffic light when traffic is stopped. Similarly, to ground an emotionally-loaded claim, you give an example of a specific situation that you'd mentally classify within that claim.

Let's look at a couple examples of people introspecting on the specific causes of their emotions.

A "Disgusting Slob" Roommate (h/t [Eliezer](#))

Tim feels strongly that his roommate Rob is a disgusting slob, so he says, "Hey Rob, can you stop being such a slob?" and Rob says "I'm not a slob, you're a neat freak!" And they start to hate each other.

Let's see how Tim can do better by activating his specificity powers!

In general, when you activate your specificity powers on an emotional claim you've made, you might end up instantly demolishing the claim, like a boy who's made to realize that there's nothing inherently "mean" about his sister giving him a soft punch on the arm. But if you find that your emotion *can* be coherently grounded in the specific stimuli you've experienced, then you can at least figure out a *specific counterfactual situation* under which your emotional response would no longer be triggered.

Let's see what happens in Tim's case.

First, Tim grounds his concept of a "disgusting slob": Tim posits that "someone who often leaves soda cans on the table, often leaves dishes in the sink, and never vacuums" is a specific stimulus that would be sufficient to mentally classify anyone as a disgusting slob. So Rob really *is* currently a disgusting slob, from the perspective of Tim's emotional processes.



But at least Tim can figure out a specific counterfactual situation under which his emotional response would no longer be triggered: If Rob were to stop leaving soda cans on the table, stop leaving dishes in the sink, and vacuum sometimes, then Rob would no longer have any of the specific characteristics that comprise Tim's emotionally-loaded concept of "slob". At that point, Tim's feeling of Rob "being a slob" will be demolished. Rob wouldn't be a slob anymore.

As an empathetic guy, Tim thinks the term "disgusting" is too harshly judgmental of poor Rob. But Tim doesn't have much control over the emotional part of his brain, the part that automatically applies abstract concept-labels like "disgusting" to stimuli before his sense of empathy gets a chance to weigh in. Tim's emotional brain automatically slides specific stimuli *up* the ladder of abstraction, and all Tim can do is painstakingly reverse-engineer the process.

An "Unfriendly" Q&A Community

When you slide any concept up or down the ladder of abstraction, you may find that the lower-level description of reality has a different qualitative character than the higher-level description. This is called "[emergence](#)". For instance, liquid water emerges from a set of non-wet H₂O molecules.

Similarly, an emotional judgment about a set of specific things may not apply to any of the things individually. We can call this **emotional emergence**. For instance, Sara Chipps, Director of Public Q&A at Stack Overflow, recently [posted](#) about a time when she felt "attacked and diminished" by her coworkers:

On a Friday afternoon, we introduced a new company-wide policy that I felt was relatively benign. What happened next was that, from my point of view, the engineering team completely lost it. No one agreed with this policy, and they made it known over seemingly hundreds of Slack pings. After an afternoon of

going back and forth, I walked away feeling emotionally drained. What had happened to my amazing coworkers that were so kind and wonderful? I felt attacked and diminished. It seemed people weren't valuing my work or my judgment.

[...] As I went back through that Friday afternoon chat log, I was shocked to see that no one had been hurling insults. There was no one saying mean things about me or attacking my efficacy directly.

[...] My coworkers hadn't become monsters, they were still the kind and caring people I thought they were. The monster in this case is not one person, it was created when lots of people, even with great intentions, publicly disagreed with you at the same time.

Sara realized that this incident is related to why new users of Stack Overflow's Q&A community often reported feeling that the community is "unfriendly". It's another example of emotional emergence:

In our developer survey results we read things like this:

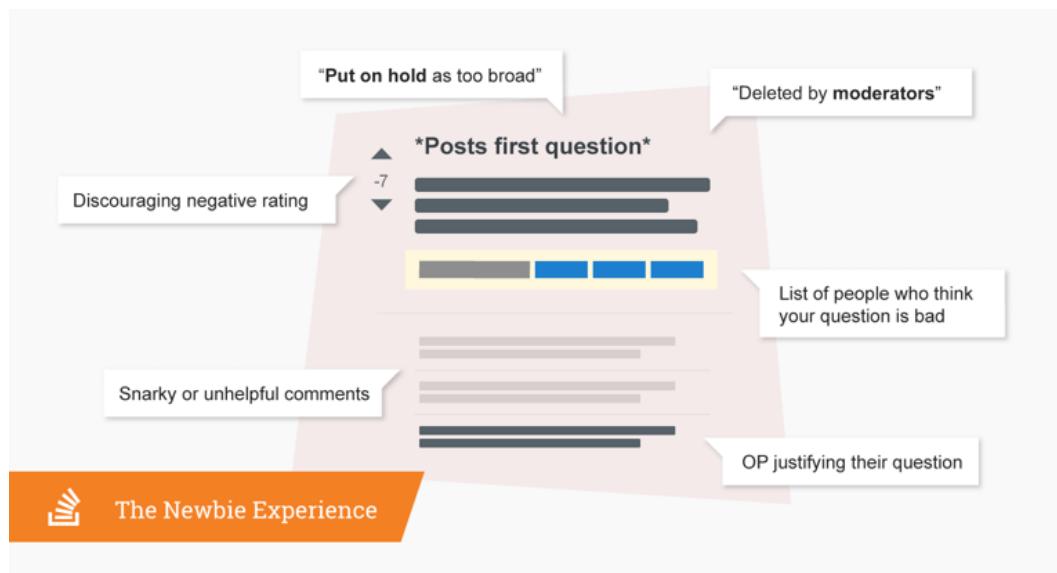
"Caustic community for new users. There is no excuse for not being kind!" [...]

"People could be less brutal" [...]

However, when our more experienced users hear this feedback they ask us to provide them with definitive examples of WHERE EXACTLY people are being unfriendly? There isn't a lot of name calling or anger, why are they being accused of being unfriendly?

When we first learn the survey participants' judgment that Stack Overflow is "unfriendly", we might mistakenly conclude that individual members of the community must not be acting friendly enough. The Stack Overflow team didn't jump to that conclusion; they kept an open mind while figuring out how to ground "unfriendly".

The team realized that the unfriendliness was triggered by how their site design collocates an overwhelming amount of negative feedback onto the same screen:



The judgment "unfriendly" turned out to be triggered by a combination of specific stimuli that were all fixable without improving the friendliness of any individual community members.

Being emotionally mature is pretty straightforward. So why aren't people more emotionally mature?

We've [seen](#) that sliding concepts down the ladder of abstraction generally goes against the grain of your typical flow of thoughts. It's hard cognitive work.

Introspecting on the specific causes of your emotions is usually even harder work, because your brain is also spinning up various automatic processes as part of the emotional reaction you're having. If you're happy, your brain spins up processes to celebrate and spread your excitement. If you're angry, your brain spins up processes to yell and stomp around and plot revenge.

By building up your specificity muscles, you'll have an easier time using them in emotionally-charged situations, so you'll be more emotionally mature.

Next post: [The Power to Teach Concepts Better](#)

Integrating the Lindy Effect

Suppose the following:

1. Your intelligence is directly proportional to how many useful things you know.
2. Your intelligence increases when you learn things and decreases as the world changes and the things you know go out-of-date.

How quickly the things you know become irrelevant is directly proportional to how many relevant things you know and therefore proportional to your intelligence I and inversely proportional to the typical lifetime of things you know L . Let's use R to denote your rate of learning. Put this together and we get a equation.

$$\frac{dt}{dt} + \frac{dt}{t} \propto R$$

If we measure intelligence in units of "facts you know" then the proportionality becomes an equality.

$$\frac{dt}{dt} = R - \frac{1}{t}$$

The solution to this first order differential equation is an exponential function.

$$I(t) = c e^{-t/L} + R L$$

We must solve for c . For convenience let's declare that your intelligence is 0 at time $t = 0$. Then c must equal $-RL$. That gives us a tidy solution.

$$I(t) = RL(1 - e^{-t/L})$$

Our solution makes sense intuitively because your intelligence is directly proportional to R and L . But wait a minute. L isn't just a coefficient. It's in the exponential too.

Time t and Lifetime L

Most human beings reading this article will be between 10 years and 100 years old. In other words, t is measured in decades. In other other words, t is on the order of 10 years.

L values, on the other hand, are distributed exponentially across many orders of magnitude.

Order of days. (0.003 years)

- daily newspaper

Order of weeks (0.2 years)

- Sunday newspaper
- political story
- sports game outcome

Order of decades (10 years)

- programming languages

Order of centuries (100 years)

- classic literature
- most spoken languages

Order of millennia (1,000 years)

- cooking
- history

Order of 10,000 years

- human psychology

Order of gigaannum (billion years)

- biology
- physics

Forever

- math

The details of whether exactly each of these things fit on the scale is not important. What is important is that most things you can know have a useful lifetime at least one order of magnitude away from the human timescale of decades. In other words, we can assume that either L is much greater than t or much less than t .

Suppose L is much less than t . Then the exponential vanishes and we're left with $I = RL$. In other words, if $L \ll t$ then how long you have been learning for t is irrelevant. I is constant with respect to time. Years and years of studying will not make you smarter over time.

Suppose that L is much greater than t . Then $I = Rt$. What used to be a constant function becomes an increasing linear function.

$$I(t) = \begin{cases} L & \text{for } t \gg L \\ R \cdot t & \text{for } t \ll L \end{cases}$$

$I = Rt$ grows with respect to time while $I = RL$ stays constant. Eventually, anyone on an $I = Rt$ trajectory will always become smarter than someone on an $I = RL$ trajectory even if the person on the $I = RL$ trajectory has higher $R \neq 0$.

In the long term, the lifetime of things you learn L is far more important than how fast you learn R . Over a lifetime of decades, someone who learns a few durable things slowly will eventually become smarter than someone who learns many transient ones quickly.

Three Stories for How AGI Comes Before FAI

Epistemic status: fake framework

To do effective differential technological development for AI safety, we'd like to know which combinations of AI insights are more likely to lead to FAI vs UFAI. This is an overarching strategic consideration which feeds into questions like how to think about the value of [AI capabilities research](#).

As far as I can tell, there are actually several different stories for how we may end up with a set of AI insights which makes UFAI more likely than FAI, and these stories aren't entirely compatible with one another.

Note: In this document, when I say "FAI", I mean any superintelligent system which does a good job of helping humans (so an "aligned Task AGI" also counts).

Story #1: The Roadblock Story

Nate Soares describes the roadblock story in [this comment](#):

...if a safety-conscious AGI team asked how we'd expect their project to fail, the two likeliest scenarios we'd point to are "your team runs into a capabilities roadblock and can't achieve AGI" or "your team runs into an alignment *roadblock* and *can easily tell that the system is currently misaligned, but can't figure out how to achieve alignment in any reasonable amount of time.*"

(emphasis mine)

The roadblock story happens if there are key safety insights that FAI needs but AGI doesn't need. In this story, the knowledge needed for FAI is a superset of the knowledge needed for AGI. If the safety insights are difficult to obtain, or no one is working to obtain them, we could find ourselves in a situation where we have all the AGI insights without having all the FAI insights.

There is subtlety here. In order to make a strong argument for the existence of insights like this, it's not enough to point to failures of existing systems, or describe hypothetical failures of future systems. You also need to explain why the insights necessary to create AGI wouldn't be sufficient to fix the problems.

Some possible ways the roadblock story could come about:

- Maybe safety insights are more or less agnostic to the chosen AGI technology and can be discovered in parallel. (Stuart Russell has pushed against this, saying that in the same way making sure bridges don't fall down is part of civil engineering, safety should be part of mainstream AI research.)
- Maybe safety insights require AGI insights as a prerequisite, leaving us in a precarious position where we will have acquired the capability to build an AGI before we begin critical FAI research.

- This could be the case if the needed safety insights are mostly about how to safely assemble AGI insights into an FAI. It's possible we could do a bit of this work in advance by developing "contingency plans" for how we would construct FAI in the event of combinations of capabilities advances that seem plausible.
 - Paul Christiano's IDA framework could be considered a contingency plan for the case where we develop much more powerful imitation learning.
 - Contingency plans could also be helpful for directing differential technological development, since we'd get a sense of the difficulty of FAI under various tech development scenarios.
- Maybe there will be multiple subsets of the insights needed for FAI which are sufficient for AGI.
 - In this case, we'd like to speed the discovery of whichever FAI insight will be discovered last.

Story #2: The Security Story

From [Security Mindset and the Logistic Success Curve](#):

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.

The main difference between the security story and the roadblock story is that in the security story, it's not obvious that the system is misaligned.

We can subdivide the security story based on the ease of fixing a flaw if we're able to detect it in advance. For example, vulnerability #1 on the [OWASP Top 10](#) is injection, which is typically easy to patch once it's discovered. Insecure systems are often right next to secure systems in program space.

If the security story is what we are worried about, it could be wise to try & develop the AI equivalent of OWASP's [Cheat Sheet Series](#), to make it easier for people to find security problems with AI systems. Of course, many items on the cheat sheet would be speculative, since AGI doesn't actually exist yet. But it could still serve as a useful starting point for brainstorming flaws.

Differential technological development could be useful in the security story if we push for the development of AI tech that is easier to secure. However, it's not clear how confident we can be in our intuitions about what will or won't be easy to secure. In his book *Thinking Fast and Slow*, Daniel Kahneman describes his adversarial collaboration with expertise researcher Gary Klein. Kahneman was an expertise skeptic, and Klein an expertise booster:

We eventually concluded that our disagreement was due in part to the fact that we had different experts in mind. Klein had spent much time with fireground commanders, clinical nurses, and other professionals who have real expertise. I

had spent more time thinking about clinicians, stock pickers, and political scientists trying to make unsupportable long-term forecasts. Not surprisingly, his default attitude was trust and respect; mine was skepticism.

...

When do judgments reflect true expertise? ... The answer comes from the two basic conditions for acquiring a skill:

- an environment that is sufficiently regular to be predictable
- an opportunity to learn these regularities through prolonged practice

In a less regular, or low-validity, environment, the heuristics of judgment are invoked. System 1 is often able to produce quick answers to difficult questions by substitution, creating coherence where there is none. The question that is answered is not the one that was intended, but the answer is produced quickly and may be sufficiently plausible to pass the lax and lenient review of System 2. You may want to forecast the commercial future of a company, for example, and believe that this is what you are judging, while in fact your evaluation is dominated by your impressions of the energy and competence of its current executives. Because substitution occurs automatically, you often do not know the origin of a judgment that you (your System 2) endorse and adopt. If it is the only one that comes to mind, it may be subjectively undistinguishable from valid judgments that you make with expert confidence. This is why subjective confidence is not a good diagnostic of accuracy: judgments that answer the wrong question can also be made with high confidence.

Our intuitions are only as good as the data we've seen. "Gathering data" for an AI security cheat sheet could be helpful for developing security intuition. But I think we should be skeptical of intuition anyway, given the speculative nature of the topic.

Story #3: The Alchemy Story

Ali Rahimi and Ben Recht [describe](#) the alchemy story in their Test-of-time award presentation at the NeurIPS machine learning conference in 2017 ([video](#)):

Batch Norm is a technique that speeds up gradient descent on deep nets. You sprinkle it between your layers and gradient descent goes faster. I think it's ok to use techniques we don't understand. I only vaguely understand how an airplane works, and I was fine taking one to this conference. But *it's always better if we build systems on top of things we do understand deeply?* This is what we know about why batch norm works well. But don't you want to understand why reducing internal covariate shift speeds up gradient descent? Don't you want to see evidence that Batch Norm reduces internal covariate shift? Don't you want to know what internal covariate shift *is*? Batch Norm has become a foundational operation for machine learning. *It works amazingly well. But we know almost nothing about it.*

(emphasis mine)

The alchemy story has similarities to both the roadblock story and the security story.

From the perspective of the roadblock story, "alchemical" insights could be viewed as insights which could be useful if we only cared about creating AGI, but are too unreliable to use in an FAI. (It's possible there are other insights which fall into the "usable for AGI but not FAI" category due to something other than their alchemical nature--if you can think of any, I'd be interested to hear.)

In some ways, alchemy could be worse than a clear roadblock. It might be that not everyone agrees whether the systems are reliable enough to form the basis of an FAI, and then we're looking at a [unilateralist's curse](#) scenario.

Just like chemistry only came after alchemy, it's possible that we'll first develop the capability to create AGI via alchemical means, and only acquire the deeper understanding necessary to create a reliable FAI later. (This is a scenario from the roadblock section, where FAI insights require AGI insights as a prerequisite.) To prevent this, we could try & deepen our understanding of components we expect to fail in subtle ways, and retard the development of components we expect to "just work" without any surprises once invented.

From the perspective of the security story, "alchemical" insights could be viewed as components which are *clearly* prone to vulnerabilities. Alchemical components could produce failures which are hard to understand or summarize, let alone fix. From a differential technological development point of view, the best approach may be to differentially advance less alchemical, more interpretable AI paradigms, developing the AI equivalent of reliable cryptographic primitives. (Note that explainability is [inferior](#) to [interpretability](#).)

Trying to create an FAI from alchemical components is obviously not the best idea. But it's not totally clear how much of a risk these components pose, because if the components don't work reliably, an AGI built from them may not work well enough to pose a threat. Such an AGI could work better over time if it's able to improve its own components. In this case, we might be able to program it so it periodically re-evaluates its training data as its components get upgraded, so its understanding of human values improves as its components improve.

Discussion Questions

- How plausible does each story seem?
- What possibilities aren't covered by the taxonomy provided?
- What distinctions does this framework fail to capture?
- Which claims are incorrect?

Divergence on Evidence Due to Differing Priors - A Political Case Study

(This uses a politically charged topic as an example, but I'm hoping that people are willing to try to understand the points made despite that. Politics is Hard Mode, and I'm hoping to stay at a lower difficulty level for now, so I've asked that comments not discuss the object level politics.)

Last week on twitter, I saw two very different takes on how the United States reacted to 9/11, and the consequences. They both reflected an update to people's views based on the data since that time, but the conclusions radically diverged. E.T. Jaynes posited that this can happen, but this is the first time I recognized it in practice so clearly, and thought it was worth noting simply as an example. Beyond that, I wanted to point out how it can be to some extent avoided.

The first was Don Moynihan, who [said](#): "It is important to remember 9/11 and the lives that were lost. Its also important to remember that period of American history fully, to understand that a terrorist attack triggered a series of catastrophic judgments by US politicians that led to the loss of more innocent lives."

The second was David French, who [said](#): " If you had told us on that day that we wouldn't endure another mass-scale attack on American soil for at least 18 more years, we would have thought you were wildly optimistic. The achievement of our military and security establishment should never be underestimated."

First, I want to note that the two views are based on different counterfactuals. Moynihan presumably assumes that the counterfactual rate of terrorist attacks had the US not gone to war in Afghanistan and Iraq to be at least relatively low. He therefore updates based on the fact that there have been very few credible attempts to mount large attacks on the US homeland, to conclude that they would be foiled by standard US intelligence sources and policing. French explicitly calls out the fact that the commonly held prior for the number of attacks that would be mounted in the wake of 9/11 was high, and asserts that this was correct but-for the military interventions the US waged. He therefore updates based on the fact that there have been very few credible attempts to mount large attacks on the US homeland, to conclude that the military interventions were successful.

Clearly, it is not the case that either person is ignoring the evidence. In this case, there are different reasons to update towards each of the models; the lack of credible attack attempts in the US contrasts with the large number in Iraq, and it's plausible that without the US wars abroad, some of that effort would have been directed at the US. On the other hand, law enforcement was very successful in detecting and stopping attacks, so it's plausible that few would have gotten through anyways. But since we can't see what would have happened has the US not gone to war (i.e. counterfactual realities are unobservable), we may be tempted to conclude that evidence is useless in the face of different prior beliefs. This isn't quite true.

If Moynihan and French had been asked in detail in 2001 what they expected in the case that the US would or would not go to war, they would be forced to confront the

ways in which their predictions failed. Perhaps their conclusions would be different - but most people don't routinely make quantifiable predictions. Their stated models are at best capable of being twisted, and if people want to believe their model, and not change their mind, not only can the invisible dragon in the garage be post-hoc determined to be permeable to flour once an annoying rationalist proposes a test of the theory, but given the flexibility that language offers, people often specify models of the world that don't require post-hoc adjustment, just defensible clarifications. So unless we're incredibly detailed in the predictions we request of people, the ability to use data to reinforce rather than revise beliefs can't be stopped.

Ideally, we'd have the ability to build a correct model, but we can't - certainly not in the space of this post, near-certainly not in a couple years of research into international relations theory, and plausibly not at all due to the paucity of evidence and the number of uncertain variables involved.

The better approach, I think, is to consider the outside view about the models. We have are two different models that are espoused by people with differing political viewpoints. Each of the models reflect a combination of motivated reasoning, selective blindness, and actual attempts to understand the world, and we're stuck uncertain which is less wrong.

But what we absolutely shouldn't do - and without explicitly trying not to, likely would do - is notice the model that we'd prefer and (perhaps subconsciously) preferentially interpret evidence as supporting it and disproving the alternatives. Especially here, where both models are simplified and wrong in many ways, my advice is to try to reason under model-uncertainty, instead of trying to reason the way we are naturally inclined to, by picking sides in a fight. Absent further plausible arguments and evidence - which exist, but themselves need to be evaluated very carefully for the same reasons - we should look at the models as both plausible.

The Power to Teach Concepts Better

This is Part VIII of the [Specificity Sequence](#)

When you teach someone a concept, you're building a structure in their mind by connecting up some of their mental concepts in a certain way. But you have to go in through their ears. It's kind of like building [this](#) ship-in-a-bottle LEGO set.



In this post, we'll visualize what's happening in a learner's brain and see how a teacher can wield their specificity powers to teach concepts better.

Mind-Hanging A Concept

Reading a startup's pitch begins as a learning exercise: learning what the startup does. In [How to Apply to Y Combinator](#), Paul Graham writes:

We have to read about 100 [applications] a day. That means a YC partner who reads your application will on average have already read 50 that day and have 50 more to go. Yours has to stand out. So you have to be exceptionally clear and concise. Whatever you have to say, give it to us right in the first sentence, in the simplest possible terms. [...]

The first question I look at is, "What is your company going to make?" This isn't the question I care most about, but I look at it first because I need **something to hang the application on in my mind**.

It's worth unpacking and visualizing this part of PG's advice, because we'll see that the power to **mind-hang** the concept you're trying to communicate is closely related to the power of specificity. Stay tuned for that.

First, one more snippet from PG:

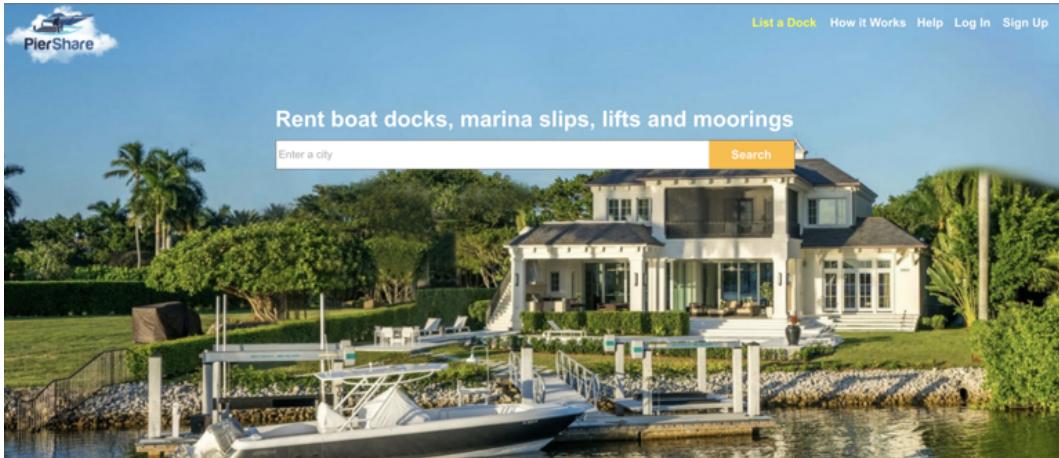
The best answers are the most matter of fact. It's a mistake to use marketing-speak to make your idea sound more exciting. We're immune to marketing-speak; to us it's just noise. So don't begin your answer with something like "*We are going to transform the relationship between individuals and information.*" That sounds impressive, but it conveys nothing. It could be a description of any technology

company. Are you going to build a search engine? Database software? A router? I have no idea. [...]

One good trick for describing a project concisely is to explain it as a variant of something the audience already knows. It's like Wikipedia, but within an organization. It's like an answering service, but for email.

I recently helped the founder of [PierShare](#) craft his YC application. For the "What is your company going to make?" question, he had originally written the first sentence:

PierShare is a new way for boat owners to easily rent a dock.



This is actually a pretty good answer because it's matter-of-fact, not marketing-speak, and I can already guess at a plausible [Value Prop Story](#). Still, I recommended taking PG's advice to describe his company more concisely as a variant of something the audience already knows:

Liron: Why not just write that you're "Airbnb for boat docks"?

Founder: Hm, how is that better though?

Liron: It builds the reader's mental model better and faster. Notice how in the next part of your application, you currently spend a lot of words describing these predictable qualities of your business:

- * You have a user-friendly online system for both parties
- * You handle the payment processing
- * You handle the insurance stuff

If you just say "Airbnb for boat docks", they instantly load a mental model of all that stuff, so you don't have to spell it out for them.

Founder: But it has important differences! Like, Airbnb's typical user just makes a one-time payment to rent someone's home for a night or two, while our typical user pays an ongoing monthly subscription to stay at someone's dock.

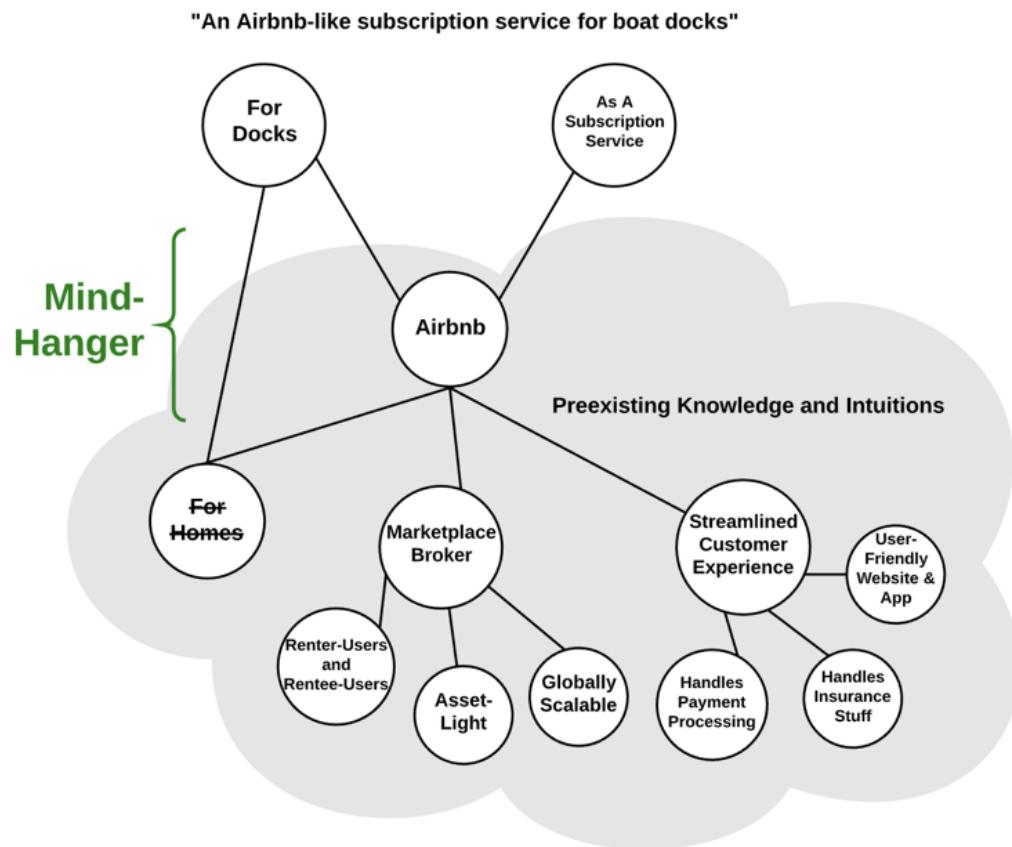
Liron: I agree that this difference is important enough to mention in your application, but still, the fastest way to build their mental model of your business is to start with their pre-built mental model of Airbnb and then apply a patch. This

is true to such a degree that if you *don't* invoke Airbnb as a shorthand to explain your business, they'll be wondering why not.

Here's what we came up with in the end:

PierShare is an Airbnb-like subscription service for boat docks.

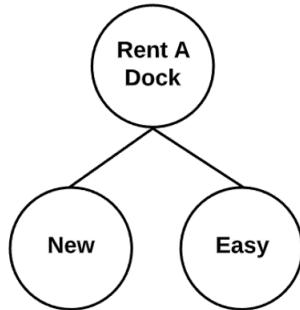
This diagram of the reader's mind illustrates how we're **mind-hanging** the concept of PierShare:



Since we know the reader is a savvy investor, it's safe to assume that the concept of "Airbnb" has an associated web of preexisting knowledge and intuitions. The reader knows that Airbnb is a marketplace broker, that marketplace brokers cater to two types of users: renters and rentees, that Airbnb has a streamlined customer experience, etc.

Compare this to "a new way for boat owners to easily rent a dock":

"A new way for boat owners to easily rent a dock"



Sure, we can expect *some* preexisting knowledge and intuitions about renting stuff in general, and boating/docks in general, but it doesn't mind-hang the way "An Airbnb-like subscription service for boat docks" does. For example, the investor won't be able to confidently predict that PierShare handles payment processing, without reading farther into the application.

Specifics Are Mind-Hangers

Here's the complete answer we crafted for "What is your company going to make?" Can you tell what trick we're using in the second and third sentences?

PierShare is an Airbnb-like service to match owners of water-docked boats (4M individuals in the US) to private dock owners (500,000 individuals in the US).

For boat owners who typically pay \$1,000–1,800/month to dock their boat (that's \$12–22k/yr, and they need this for their boat's 10- to 15-year lifetime), we let them find a privately-owned dock and save 30% (\$4k/yr).

For dock owners (people who live adjacent to a waterway) who typically leave their docks empty because it's too much friction and hassle to rent it, we give them an easy \$1k+/month revenue stream.

That's right, we're hitting them with two Value Prop Stories! (Notice that a two-sided marketplace business always needs two Value Prop Stories: in this case, one for boat owners and one for dock owners.)

One reason for using Value Prop Stories here is that we want to show off the strength of the value-delta for both sides of the marketplace: putting an extra \$4k/year into the pocket of each boat renter and \$1k/month into the pocket of each dock owner.

But the other reason we're using Value Prop Stories is simply because they add specificity to the description of what the startup does:

- Dock owners are people who live adjacent to a waterway
- Boat owners pay a lot for docking
- Every year, there's a lot of money at stake for both

We're using Value Prop Stories to slide "Airbnb for boat docks" down the ladder of abstraction.

And in general, any time you're trying to explain a concept to someone—in this case, the concept of PierShare—it's helpful to slide it down the ladder of abstraction because specific details function as mind-hangers.

Teach With Examples First

When you want to teach someone an abstract concept, keep in mind that people love climbing the [ladder of abstraction](#) from bottom to top. Brains are good at generalizing from specific experiences, but bad at grasping general concepts directly from generally-worded statements.

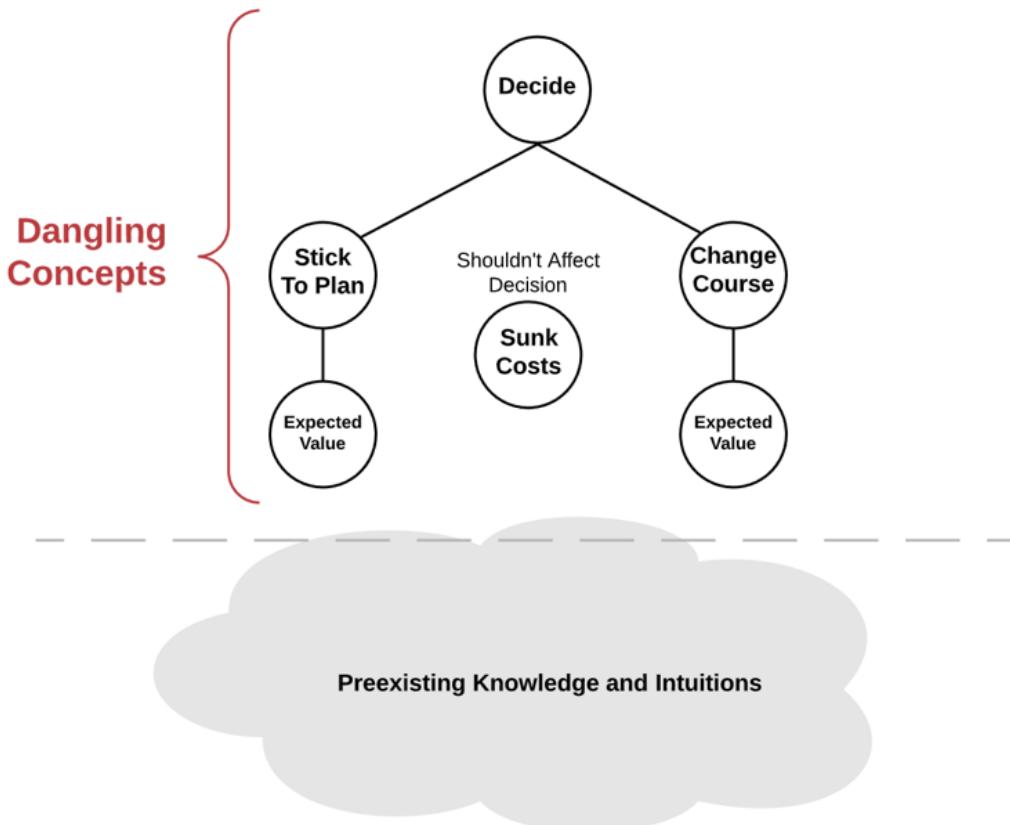
Here, let's ask your brain to grasp the following generally-worded statement of an abstract concept:

Don't Fall for the Sunk Cost Fallacy: General Statement

When deciding whether to stick to a plan or change course, your current plan's sunk costs shouldn't affect your decision, as long as you compare the total expected value of each option.

Did you manage to load the general principle into your head? Unless you immediately recognized this as a familiar principle, you probably had to read the description multiple times, slowly.

As your author, I should try not to make your brain do this! It's bad explanatory writing to hit you with a general principle without first giving you a mind-hanger for it. Look at this mess of dangling concepts I tried to fling at you:



While the concepts I tried to fling at you have some internal structure to hold them together, they're currently just **dangling** over your brain's huge web of preexisting knowledge and intuitions. Frustrating, right?

But you probably won't even think to blame me for putting you in this situation. You'll probably blame yourself for being too "dumb" to deal with the dangling concepts that the "smart" author is trying to teach you.

Why did I put you in this compromised position? I guess I expected that you'd just keep my general statement dangling in your precious working memory, and keep reading onward to the next thing I'm going to say.

But now that we have the language of *mind-hanging* and *dangling*, we can see that this is an authorially-rude thing for me to do:

Hey chump, just keep reading my explanation while your brain thrashes on my dangling concepts. I have no problem hanging this in *my* mind, so what do I care? I'll give you some mind-hangers whenever I feel like it.

Alright, let's see the kind of mind-hanger I should have given you.

Don't Fall for the Sunk Cost Fallacy: Specific Example

Imagine you've paid \$50 for tickets to see your favorite band perform a concert, but on the day of the concert, an unexpected blizzard hits your town. Now you have to decide whether to make the stressful one-hour drive through the blizzard to see the concert.

First, let's assign dollar values to all your options (approximating your [utility](#) function):

We'll define **\$0** to be the baseline value of your average evening when you haven't bought concert tickets.

Seeing the concert is worth **+\$80** to you.

(In other words, you were happy to buy the ticket for \$50, but you wouldn't have bought it for more than \$80.)

Driving through the blizzard is worth **-\$100** to you.

Now all you have to do is compare the expected value of your two options: You can drive to the concert for an expected value of **(-\$50 ticket) + (-\$100 blizzard)** + **(\$80 concert)** = **-\$70**, or do something else for an expected value of **(-\$50 ticket) + (\$0 average evening)** = **-\$50**.

Since **-\$70 < -\$50**, you should forget about the concert. It's not worth going because of the blizzard. But if you're not trained in decision theory, your intuitive thought process might go like this:

*"The concert is worth **+\$80**, and I've paid \$50 for the tickets, so if I don't drive to the concert, I'm missing out on the value of the concert and the value of the tickets, which is \$130 of value!"*

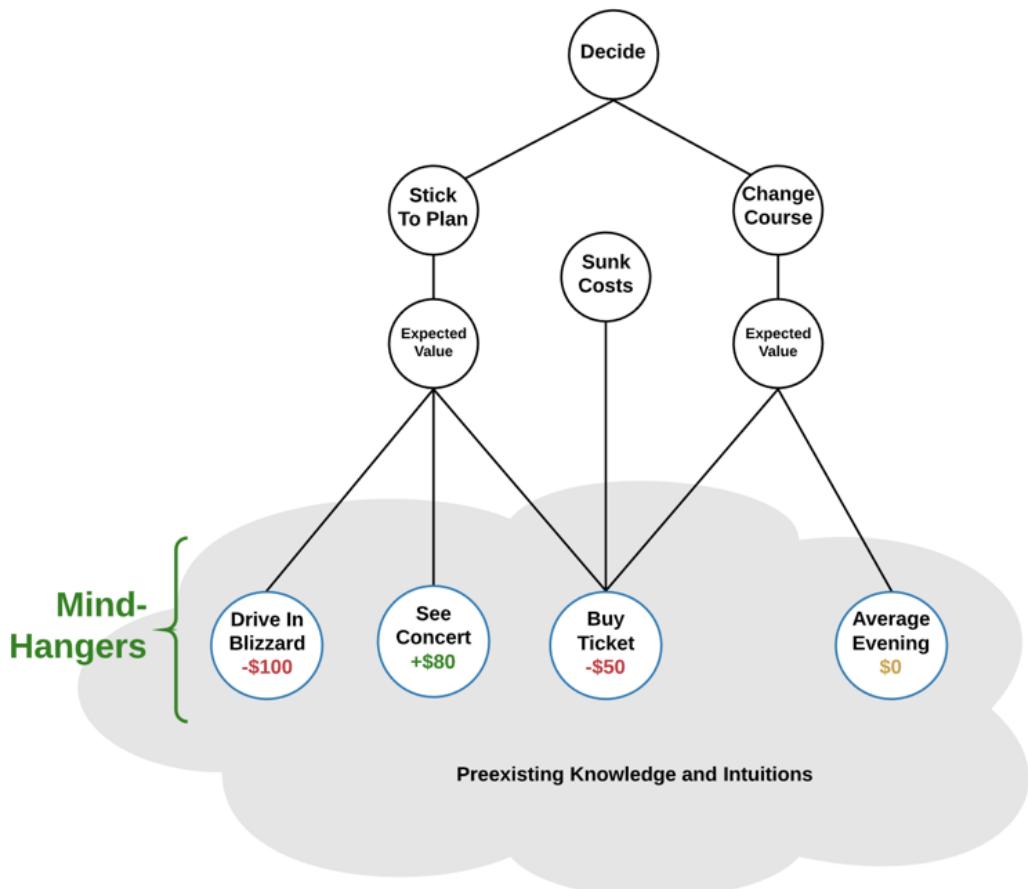
Here we say that your intuition is committing the sunk cost fallacy, because the non-refundable ticket price (\$50) is a sunk cost which has no bearing on how much value you'll gain or lose via your current decision whether to drive to the concert.

Hopefully that example was pretty easy to follow. And now that you've loaded it into your mind, you can use it as a mind-hanger to understand the general statement I'm trying to teach you:

Don't Fall for the Sunk Cost Fallacy: General Statement

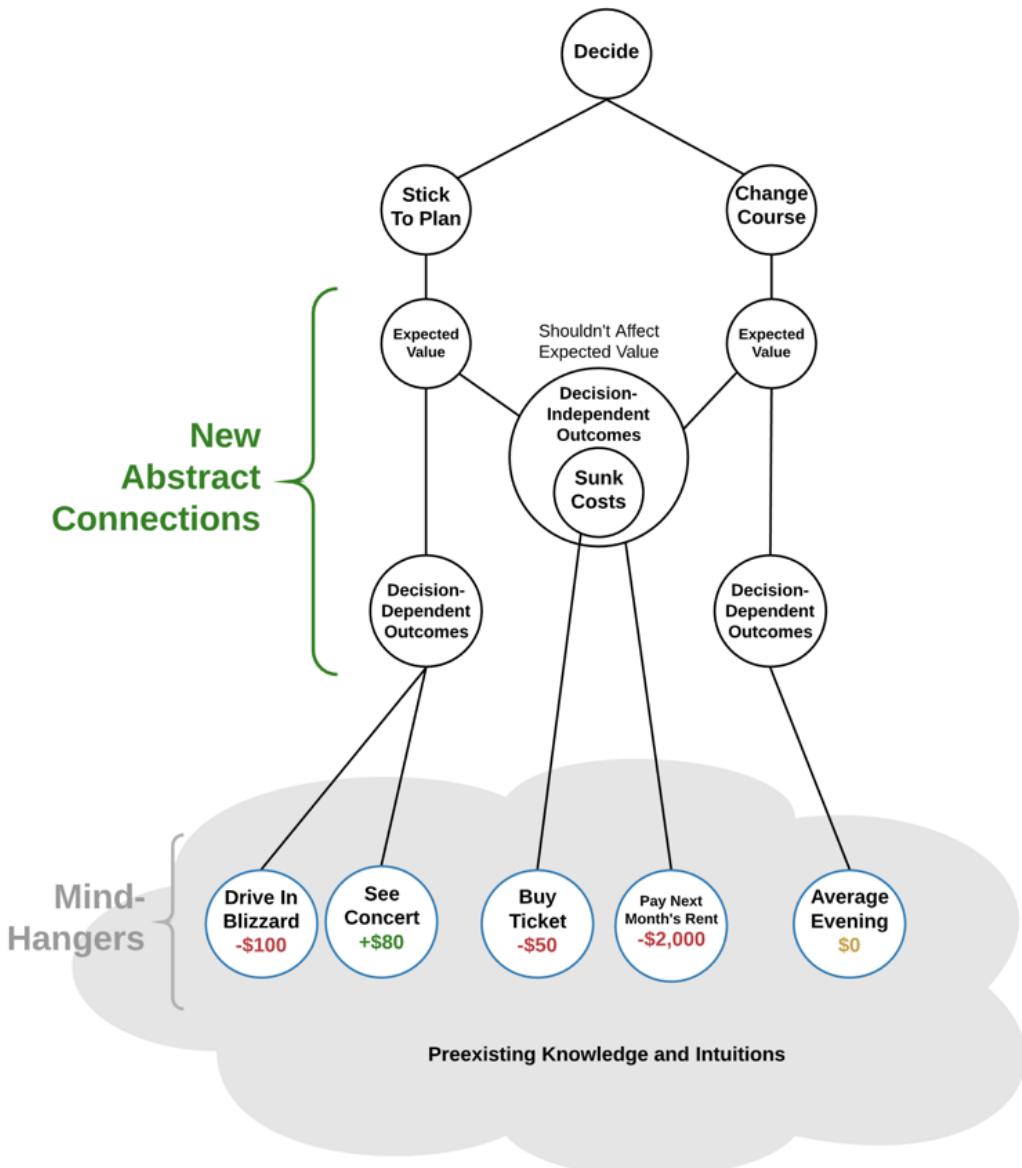
When deciding whether to stick to a plan or change course, your current plan's sunk costs shouldn't affect your decision, as long as you compare the total expected value of each option.

The general statement probably feels more understandable now that you've installed mind-hangers for it.



My goal here obviously wasn't to teach you about the [sunk cost fallacy](#), it was to show you what it feels like when an author forces you to read a general description of a concept instead of spoon-feeding you specific details that build on your preexisting knowledge and intuitions. It's easier to load a generic concept into your head if you first install mind-hangers for it.

By the way, my specific story even helped me understand the general concept better, because I realized that the **-\$50** of buying the ticket appears in both Expected Value terms, so it cancels out of the expected-value comparison. So I was able to connect my abstract concept of "sunk costs" more directly to my abstract concept of expected value:



The mind-hanging power of specific examples helps you teach others better and teach yourself better. Good to know, right?

In 2007, I stumbled on a life-changing blog post: [My favorite pedagogical principle: examples first!](#) by the mathematician Tim Gowers. It has transformed how I've been communicating concepts ever since. It was also the first inkling I ever got about the power of specificity. (It was a few more years before the dam broke and the other powers came rushing into my purview.)

Here's what Tim says is "a very simple idea that can dramatically improve the readability of just about anything":

Present examples before you discuss general concepts.

So I'm not making an original point here; I'm repeating Gowers's point, albeit with a bit less math. I'm also classifying "teach with examples first" as a kind of specificity-

based mind-hanging power that deserves to be collected and showcased together with other specificity powers in a stupendous listicle.

The Tragedy of Human Working Memory

Why does our reading or listening comprehension break down when we're presented with a general concept before a specific example?

When you were first reading my 32-word general statement about avoiding the sunk cost fallacy, it taxed your working memory more than usual, and probably overloaded it. On the other hand, when you were reading my example of driving through a blizzard to see the concert that you bought tickets for, it didn't demand as much of your working memory, even though the total number of words was almost 10x higher. Why is that?

One of the saddest things about the human condition is how underpowered your working memory is. Your brain has 100B neurons and is the smartest thing in the known universe, but remembering a seven-digit phone number is a struggle—are you kidding me?

You can be a highly intelligent person, but if I fling a handful of dangling concepts at you and you're not ready with a sturdy mind-hanger, you'll just try to fasten them down with whatever you can: a few measly strips of working-memory duct tape.



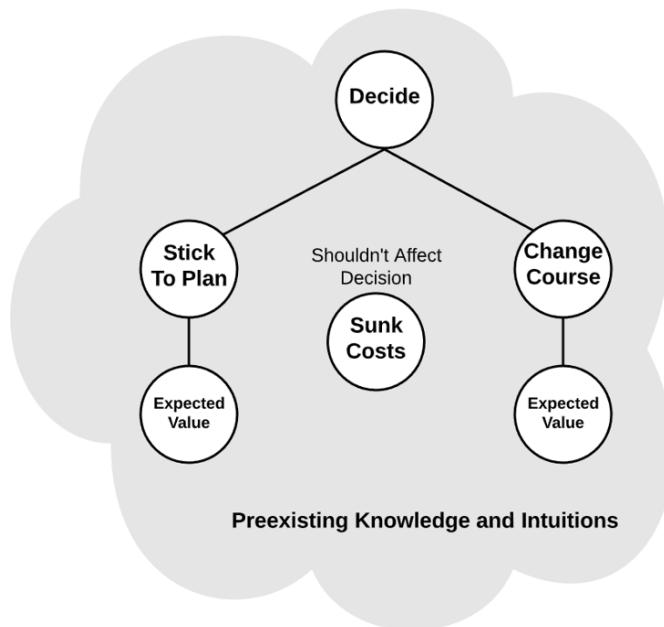
It's sad because the amount of working memory we have might be an easily-tweakable genetic parameter. Natural Selection might have been stingy with this parameter's value because we had enough for our ancestors' needs. But now we're trying to build a technological society and it's clearly not enough working memory for *our* needs.

A superintelligent mind with a reasonable amount of working memory could process generic statements all day long and never whine about dangling concepts. (I feel like the really smart people on LessWrong and Math Overflow also exhibit this behavior to some degree.) But as humans with tragically limited short-term memories, we need all the help we can get. We need our authors and teachers to give us mind-hangers.

Mind-Hanging vs. Grounding

We've [defined](#) **grounding** a term as describing a specific stimulus that you'd mentally classify within that term. That kind of description falls within the shared space of both conversation participants' preexisting knowledge and intuitions. So any grounding is also a valid mind-hanging.

On the other hand, you're allowed to mind-hang a concept on any other concept that your conversation partner is familiar with, so it's sometimes possible to offer people mind-hangings that aren't also groundings. For example, if our conversation partner was extremely familiar with decision theory, our general statement about the sunk cost fallacy wouldn't be dangling concept in their mind; the whole structure could hang entirely within their preexisting knowledge and intuitions.



Next post: [Is Specificity a Mental Model?](#)

On Becoming Clueless

It is said that every year the IQ needed to destroy the world drops by one point.

Well, yes, but let me add a different spin on the problem:

Every year, the IQ needed to make sense of the world raises by one point.

If your IQ is 100 and you want to see yourself in 2039 just ask somebody with IQ 80 and listen carefully.

I know that some people are troubled about prospects of those less intellectually gifted in the modern knowledge-based economy. And yes, it's troubling that we are heading towards some kind of intellectual elitism. But, on the other hand, it may be just a temporary thing. At the end we will all, village idiots and von Neumanns alike, end up having no clue.

What are we going to do then?

The strategy-stealing assumption

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Suppose that 1% of the world's resources are controlled by unaligned AI, and 99% of the world's resources are controlled by humans. We might hope that at least 99% of the universe's resources end up being used for stuff-humans-like (in expectation).

Jessica Taylor argued for this conclusion in [Strategies for Coalitions in Unit-Sum Games](#): if the humans divide into 99 groups each of which acquires influence as effectively as the unaligned AI, then by symmetry each group should end up with as much influence as the AI, i.e. they should end up with 99% of the influence.

This argument rests on what I'll call the *strategy-stealing assumption*: for any strategy an unaligned AI could use to influence the long-run future, there is an analogous strategy that a similarly-sized group of humans can use in order to capture a similar amount of flexible influence over the future. By "flexible" I mean that humans can decide later what to do with that influence—which is important since humans don't yet know what we want in the long run.

Why might the strategy-stealing assumption be true?

Today there are a bunch of humans, with different preferences and different kinds of influence. Crudely speaking, the long-term outcome seems to be determined by some combination of {which preferences have how much influence?} and {what is the space of realizable outcomes?}.

I expect this to become more true over time—I expect groups of agents with diverse preferences to eventually approach efficient outcomes, since otherwise there are changes that every agent would prefer (though this is not obvious, especially in light of bargaining failures). Then the question is just about *which* of these efficient outcomes we pick.

I think that our actions don't effect the space of realizable outcomes, because long-term realizability is mostly determined by facts about distant stars that we can't yet influence. The obvious exception is that if we colonize space faster, we will have access more resources. But [quantitatively this doesn't seem like a big consideration](#), because astronomical events occur over millions of millennia while our decisions only change colonization timelines by decades.

So I think our decisions mostly affect long-term outcomes by changing the relative weights of different possible preferences (or by causing extinction).

Today, one of the main ways that preferences have weight is because agents with those preferences control resources and other forms of influence. Strategy-stealing seems most possible for this kind of plan—an aligned AI can exactly copy the strategy of an unaligned AI, except the money goes into the aligned AI's bank account instead. The same seems true for most kinds of resource gathering.

There are lots of strategies that give influence to other people instead of helping me. For example, I might preferentially collaborate with people who share my values. But I

can still steal these strategies, as long as my values are just as common as the values of the person I'm trying to steal from. So a majority can steal strategies from a minority, but not the other way around.

There can be plenty of strategies that don't involve acquiring resources or flexible influence. For example, we could have a parliament with obscure rules in which I can make maneuvers that advantage one set of values or another in a way that can't be stolen. Strategy-stealing may only be possible at the level of groups—you need to retain the option of setting up a different parliamentary system that doesn't favor particular values. Even then, it's unclear whether strategy-stealing is possible.

There isn't a clean argument for strategy-stealing, but I think it seems plausible enough that it's meaningful and productive to think of it as a plausible default, and to look at ways it can fail. (If you found enough ways it could fail, you might eventually stop thinking of it as a default.)

Eleven ways the strategy-stealing assumption could fail

In this section I'll describe some of the failures that seem most important to me, with a focus on the ones that would interfere with the argument in the introduction.

1. AI alignment

If we can build smart AIs, but not aligned AIs, then humans can't necessarily use AI to capture flexible influence. I think this is the most important way in which strategy-stealing is likely to fail. I'm not going to spend much time talking about it here because I've spent so much time elsewhere.

For example, if smart AIs inevitably want to fill the universe with paperclips, then "build a really smart AI" is a good strategy for someone who wants to fill the universe with paperclips, but it can't be easily stolen by someone who wants anything else.

2. Value drift over generations

The values of 21st century humans are determined by some complicated mix of human nature and the modern environment. If I'm a 16th century noble who has really specific preferences about the future, it's not really clear how I can act on those values. But if I'm a 16th century noble who thinks that future generations will inevitably be wiser and should get what they want, then I'm in luck, all I need to do is wait and make sure our civilization doesn't do anything rash. And if I have some kind of crude intermediate preferences, then I might be able to push our culture in appropriate directions or encourage people with similar genetic dispositions to have more kids.

This is the most obvious and important way that strategy-stealing has failed historically. It's not something I personally worry about too much though.

The big reason I don't worry is some combination of common-sense morality and decision-theory: our values are the product of many generations each giving way to the next one, and so I'm pretty inclined to "pay it forward." Put a different way, I think it's relatively clear I should empathize with the next generation since I might well have been in their place (whereas [I find it much less clear under what conditions I should empathize with AI](#)). Or from yet another perspective, the same intuition that I'm

“more right” than previous generations makes me very open to the possibility that future generations are more right still. This question gets very complex, but my first-pass take is that I’m maybe an order of magnitude less worried than about other kinds of value drift.

The small reason I don’t worry is that I think this dynamic is probably going to be less important in the future (unless we actively want it to be important—which seems quite possible). I believe there is a good chance that within 60 years most decisions will be made by machines, and so the handover from one generation to the next will be optional.

That all said, I am somewhat worried about more “out of distribution” changes to the values of future generations, in scenarios where AI development is slower than I expect. For example, I think it’s possible that genetic engineering of humans will substantially change what we want, and that I should be less excited about that kind of drift. Or I can imagine the interaction between technology and culture causing similarly alien changes. These questions are even harder to think about than the basic question of “how much should I empathize with future generations?” which already seemed quite thorny, and I don’t really know what I’d conclude if I spent a long time thinking. But at any rate, these things are not at the top of my priority queue.

3. Other alignment problems

Alts and future generations aren’t the only optimizers around. For example, we can also build institutions that further their own agendas. We can then face a problem analogous to AI alignment—if it’s easier to build effective institutions with some kinds of values than others, then those values could be at a structural advantage. For example, we might inevitably end up with a society that optimizes generalizations of short-term metrics, if big groups of humans are much more effective when doing this. (I say “generalizations of short-term metrics” because an exclusive focus on short-term metrics is the kind of problem that can fix itself over the very long run.)

I think that institutions are currently considerably weaker than humans (in the sense that’s relevant to strategy-stealing) and this will probably remain true over the medium term. For example:

- A company with 10,000 people might be much smarter than any individual humans, but mostly that’s because of its alliance with its employees and shareholders—most of its influence is just used to accumulate more wages and dividends. Companies do things that seem antisocial not because they have come unmoored from any human’s values, but because plenty of influential humans want them to do that in order to make more money. (You could try to point the “market” as an organization with its own preferences, but it’s even worse at defending itself than bureaucracies—it’s up to humans who benefit from the market to defend it.)
- Bureaucracies can seem unmoored from any individual human desire. But their actual ability to defend themselves and acquire resources seems much weaker than other optimizers like humans or corporations.

Overall I’m less concerned about this than AI alignment, but I do think it is a real problem. I’m somewhat optimistic that the same general principles will be relevant both to aligning institutions and Alts. If AI alignment wasn’t an issue, I’d be more concerned by problems like institutional alignment.

4. Human fragility

If AI systems are aligned with humans, they may want to keep humans alive. Not only do humans prefer being alive, humans may need to survive if they want to have the time and space to figure out what they really want and to tell their AI what to do. (I say “may” because at some point you might imagine e.g. putting some humans in cold storage, to be revived later.)

This could introduce an asymmetry: an AI that just cares about paperclips can get a leg up on humans by threatening to release an engineered plague, or trashing natural ecosystems that humans rely on. (Of course, this asymmetry may also go the other way—values implemented in machines are reliant on a bunch of complex infrastructure which may be more or less of a liability than humanity’s reliance on ecosystems.)

Stepping back, I think the fundamental long-term problem here is that “do what this human wants” is only a simple description of human values if you actually have the human in hand, and so an agent with these values does have a big extra liability.

I do think that the extreme option of “storing” humans to revive them later is workable, though most people would be very unhappy with a world where that becomes necessary. (To be clear, I think it almost certainly won’t.) We’ll return to this under “short-term terminal preferences” below.

5. Persuasion as fragility

If an aligned AI defines its values with reference to “whatever Paul wants,” then someone doesn’t need to kill Paul to mess with the AI, they just need to change what Paul wants. If it’s very easy to manipulate humans, but we want to keep talking with each other and interacting with the world despite the risk, then this extra attack surface could become a huge liability.

This is easier to defend against—just stop talking with people except in extremely controlled environments where you can minimize the risk of manipulation—but again humans may not be willing to pay that cost.

The main reason this might be worse than point 4 is that humans may be relatively happy to physically isolate themselves from anything scary, but it would be much more costly for us to cut off from contact with other humans.

6. Asymmetric persuasion

Even if humans are the only optimizers around, it might be easier to persuade humans of some things than others. For example, you could imagine a world where it’s easier to convince humans to endorse a simple ideology like “maximize the complexity of the universe” than to convince humans to pursue some more complex and subtle values.

This means that people with easily-persuadable values can use persuasion as a strategy, and people with other values cannot copy it.

I think this is ultimately more important than fragility, because it is relevant before we have powerful AI systems. It has many similarities to “value drift over generations,” and I have some mixed feelings here as well—there are some kinds of argument and

deliberation that I certainly do endorse, and to the extent that my current views are the product of significant amounts of non-endorsed deliberation I am more inclined to be empathetic to future people who are influenced by increasingly-sophisticated arguments.

But as I described in section 2, I think these connections can get weaker as technological progress moves us further out of distribution, and if you told me that e.g. it was possible to perform a brute force search and find an argument that could convince someone to maximize the complexity of the future, I wouldn't conclude that it's probably fine if they decided to do that.

(Credit to Wei Dai for emphasizing this failure mode.)

7. Value-sensitive bargaining

If a bunch of powerful agents collectively decide what to do with the universe, I think it probably won't look like "they all control their own slice of the universe and make independent decisions about what to do." There will likely be opportunities for trade, they may have meddling preferences (where I care what you do with your part of the universe), there may be a possibility of destructive conflict, or it may look completely different in an unanticipated way.

In many of these settings the outcome is influenced by a complicated bargaining game, and it's unclear whether the majority can steal a minority's strategy. For example, suppose that there are two values X and Y in the world, with 99% X-agents and 1% Y-agents. The Y-agents may be able to threaten to destroy the world unless there is an even split, and the X-agents have no way to copy such a strategy. (This could also occur over the short term.)

I don't have a strong view about the severity of this problem. I could imagine it being a big deal.

8. Recklessness

Some preferences might not care about whether the world is destroyed, and therefore have access to productive but risky strategies that more cautious agents cannot copy. The same could happen with other kinds of risks, like commitments that are game-theoretically useful but risk sacrificing some part of the universe or creating long-term negative outcomes.

I tend to think about this problem in the context of particular technologies that pose an extinction risk, but it's worth keeping in mind that it can be compounded by the existence of more reckless agents.

Overall I think this isn't a big deal, because it seems much easier to cause extinction by trying to kill everyone than as an accident. There are fewer people who are in fact trying to kill everyone, but I think not enough fewer to tip the balance. (This is a contingent fact about technology though; it could change in the future and I could easily be wrong even today.)

9. Short-term unity and coordination

Some actors may have long-term values that are easier to talk about, represent formally, or reason about. Relative to humans, AIs may be especially likely to have such values. These actors could have an easier time coordinating, e.g. by pursuing some explicit compromise between their values (rather than being forced to find a governance mechanism for some resources produced by a joint venture).

This could leave us in a place where e.g. an unaligned AI controls 1% resources, but the majority of resources are controlled by humans who want to acquire flexible resources. Then the unaligned AIs can form a coalition which achieves very high efficiencies, while the humans cannot form 99 other coalitions to compete.

This could theoretically be a problem without AI, e.g. a large group of human with shared explicit values might be able to coordinate better and so leave normal humans at a disadvantage, though I think this is relatively unlikely as a major force in the world.

The seriousness of this problem is bounded by both the efficiency gains for a large coalition, and the quality of governance mechanisms for different actors who want to acquire flexible resources. I think we have OK solutions for coordination between people who want flexible influence, such that I don't think this will be a big problem:

- The humans can participate in lotteries to concentrate influence. Or you can gather resources to be used for a lottery in the future, while still allowing time for people to become wiser and then make bargains about what to do with the universe before they know who wins.
- You can divide up the resources produced by a coalition equitably (and then negotiate about what to do with them).
- You can modify other mechanisms by allowing votes that could e.g. overrule certain uses of resources. You could have more complex governance mechanisms, can delegate different kinds of authority to different systems, can rely on trusted parties, etc.
- Many of these procedures work much better amongst groups of humans who expect to have relatively similar preferences or have a reasonable level of trust for other participants to do something basically cooperative and friendly (rather than e.g. demanding concessions so that they don't do something terrible with their share of the universe or if they win the eventual lottery).

(Credit to Wei Dai for describing and emphasizing this failure mode.)

10. Weird stuff with simulations

I think civilizations like ours mostly have an impact via the common-sense channel where we ultimately colonize space. But there may be many civilizations like ours in simulations of various kinds, and influencing the results of those simulations could also be an important part of what we do. In that case, I don't have any particular reason to think strategy-stealing breaks down but I think stuff could be very weird and I have only a weak sense of how this influences optimal strategies.

Overall I don't think much about this since it doesn't seem likely to be a large part of our influence and it doesn't break strategy-stealing in an obvious way. But I think it's worth having in mind.

11. Other preferences

People care about lots of stuff other than their influence over the long-term future. If 1% of the world is unaligned AI and 99% of the world is humans, but the AI spends all of its resources on influencing the future while the humans only spend one tenth, it wouldn't be too surprising if the AI ended up with 10% of the influence rather than 1%. This can matter in lots of ways other than literal spending and saving: someone who only cared about the future might make different tradeoffs, might be willing to defend themselves at the cost of short-term value (see sections 4 and 5 above), might pursue more ruthless strategies for expansion, and so on.

I think the simplest approximation is to restrict attention to the part of our preferences that is about the long-term (I discussed this a bit in [Why might the future be good?](#)). To the extent that someone cares about the long-term less than the average actor, they will represent a smaller fraction of this “long-term preferences” mixture. This may give unaligned AI systems a one-time advantage for influencing the long-term future (if they care more about it) but doesn't change the basic dynamics of strategy-stealing. Even this advantage might be clawed back by a majority (e.g. by taxing savers).

There are a few places where this picture seems a little bit less crisp:

- Rather than being able to spend resources on either the short or long-term, sometimes you might have preferences about *how* you acquire resources in the short-term; an agent without such scruples could potentially pull ahead. If these preferences are strong, it probably violates strategy-stealing unless the majority can agree to crush anyone unscrupulous.
- For humans in particular, it may be hard to separate out “humans as repository of values” from “humans as an object of preferences,” and this may make it harder for us to defend ourselves (as discussed in sections 4 and 5).

I mostly think these complexities won't be a big deal quantitatively, because I think our short-term preferences will mostly be compatible with defense and resource acquisition. But I'm not confident about that.

Conclusion

I think strategy-stealing isn't really true; but I think it's a good enough approximation that we can basically act as if it's true, and then think about the risk posed by possible failures of strategy-stealing.

I think this is especially important for thinking about AI alignment, because it lets us formalize the lowered goalposts I discussed [here](#): we just want to ensure that AI is compatible with strategy-stealing. These lowered goalposts are an important part of why I think we can solve alignment.

In practice I think that a large coalition of humans isn't reduced to strategy-stealing—a majority can simply stop a minority from doing something bad, rather than by copying it. The possible failures in this post could potentially be addressed by either a technical solution or some kind of coordination.

Candy for Nets

Yesterday morning my five-year-old daughter was asking me about mosquitos, and we got on to talking about malaria, [nets](#), and how [Julia](#) and I [donate](#) to the [AMF](#) to keep other kids from getting sick and potentially dying. Lily took it very seriously, and proposed that when I retire she take my [programming.job](#) and donate in my place.

I told her that she didn't need to wait until after I retired to start helping, and she decided she wanted to sell candy on the [bike path](#) as a fundraiser. I told her we could do this after naps if the weather was still nice, and the first thing she said when I got her up from her nap was that she wanted to go make a sign.

She dictated to me, "Lily is selling candy to raise money for malaria nets, \$1" and I wrote the letters. She colored them in:



(It looks like she's posing with the sign here, but this is just how she happened to position herself for coloring. She has short arms.)

Once Anna was up from her (longer) nap I got out the [wagon](#) and brought them over to the bike path. Lily did all the selling; I just hung out to the side, leaning against a tree.



She's always been good at talking to adults, and did a good job selling the candy. She would explain that the candy was \$1/each, that the money was going to buy malaria nets, and that malaria was a very bad disease that you got from mosquitoes. People were generous, and several people gave without taking candy, or put in an extra dollar. One person didn't have cash but wanted to give enough that they went home and came back with a dollar. As someone who grew up in a part of town with very little foot traffic, the idea that you can just walk a short distance from your house to somewhere where several people will pass per minute continually amazes me.

After about twenty minutes all the candy was sold and Lily had collected \$20.75. She played in the park for a while, and then when we came home she asked how we would use the money to buy nets. I showed her pictures of distributions on the [AMF website](#) but she wanted to see pictures of the nets in use so we spent a while on [image search](#):



I explained that we weren't going to distribute the nets ourselves, but that we would provide the money so other people could.

Initially she didn't want to donate the whole amount, but wanted to set aside half to buy more candy so she could do this again. I told her that I would be happy to buy the candy. Possibly I should have let her manage this herself, but I was worried that the money wouldn't end up donated which wouldn't have been fair to the people who'd bought the candy, and explained this to her. She gave me the \$20.75 and I used my credit card to pay for the nets. [1]

Here's the message she dictated for the [donation](#):

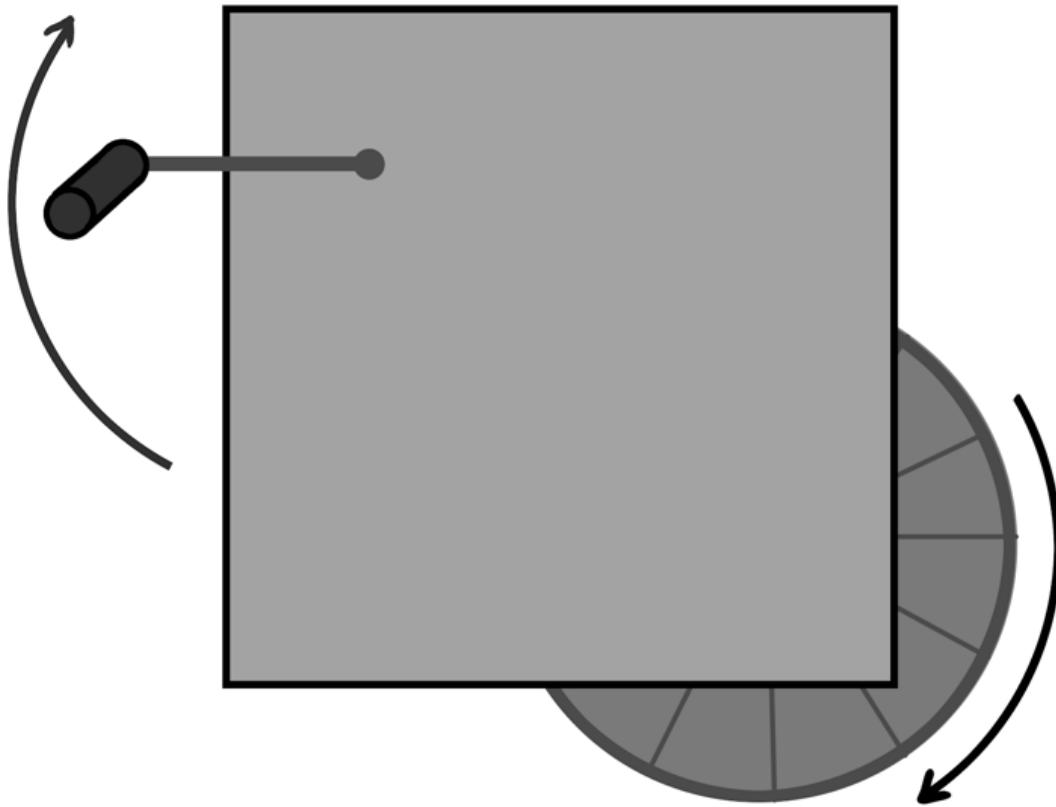
I want people to be safe in the world from biting mosquitoes. I don't want them getting hurt, and especially I don't want the kids like me to die.

I don't know how her relationship with altruism will change as she gets older, and I do think there are ways it will be hard for her to have parents who have strong unusual views. As we go I'm going to continue to try very hard not to pressure or manipulate her, while still giving advice and helping her explore her motivations here. I am, however, very proud of her today.

[1] I haven't listed this on our [donations page](#) and it doesn't count it towards our 50% goal because the donation was Lily's and not ours.

Gears vs Behavior

Thankyou to [Sisi Cheng](#) (of the [Working as Intended](#) comic) for the excellent drawings.



Suppose we have a gearbox. On one side is a crank, on the other side is a wheel which spins when the crank is turned. We want to predict the rotation of the wheel given the rotation of the crank, so we run a [Kaggle competition](#).

We collect hundreds of thousands of data points on crank rotation and wheel rotation. 70% are used as training data, the other 30% set aside as test data and kept under lock and key in an old nuclear bunker. Hundreds of teams submit algorithms to predict wheel rotation from crank rotation. Several top teams combine their models into one gradient-boosted deep random neural support vector forest. The model achieves stunning precision and accuracy in predicting wheel rotation.

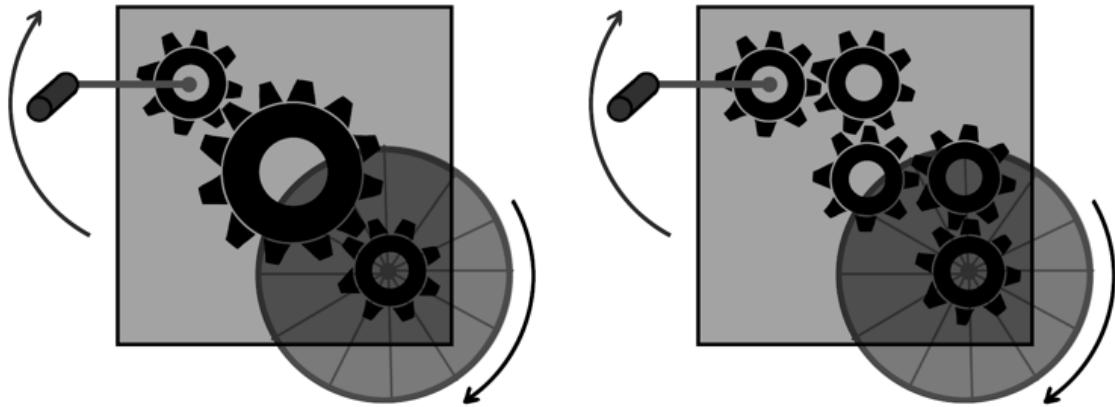
On the other hand, in a very literal sense, the model contains no gears. Is that a problem? If so, when and why would it be a problem?

What is Missing?

When we say the model “[contains no gears](#)”, what does that mean, in a less literal and more generalizable sense?

Simplest answer: the deep random neural support vector forest model does not tell us what we expect to see if we open up the physical gearbox.

For instance, consider these two gearboxes:



Both produce the same input-output behavior. Our model above, which treats the gearbox as a literal black box, does not tell us anything at all which would distinguish between these two cases. It only talks about input-output behavior, without making any predictions about what's inside the gearbox (other than that the gearbox must be consistent with the input/output behavior).

That's the key feature of gears-level models: they make falsifiable predictions about the internals of a system, separate from the externally-visible behavior. If a model could correctly predict all of a system's externally-visible behavior, but still be falsified by looking inside the box, then that's a gears-level model. Conversely, we cannot fully learn gears-level models by looking only at externally-visible input-output behavior - external behavior cannot, for example, distinguish between the 3- and 5-gear models above. A model which can be fully learned from system behavior, without any side information, is not a full gears-level model.

Why would this be useful, if what we really care about is the externally-visible behavior? Several reasons:

- First and foremost, if we are able to actually look inside the box, then that provides a huge amount of information about the behavior. If we can see the physical gears, then we can immediately make highly confident predictions about system behavior.
- More generally, any information about the internals of the system provide a "side channel" for testing gears-level models. If data about externally-visible behavior is limited, then the ability to leverage data about system internals can be valuable.
- It may be that all of our input data is only from within a certain range - i.e. we never tried cranking the box faster than a human could crank. If someone comes along and attaches a motor to the crank, then that's going to generate input way outside the range of what our input/output model has ever seen - but if we know what the gears look like, then that won't be a problem. In other words, knowing what the system internals look like lets us deal with distribution shifts.
- Finally, if someone changes something about the system, then a model trained only on input/output data will fail completely. For instance, maybe there's a switch on top of the gearbox which disconnects the gears, and nobody has ever thrown it before. If we know what the inside of the box looks like, then that's not a problem - we can look at what the switch does.

All that said, if we have abundant data and aren't worried about distribution shifts or system changes, non-gears models can still give us great predictive power. [Solomonoff induction](#) is

the idealized theoretical example: it gives asymptotically optimal predictions based on input-output behavior, without any visibility into the system internals.

Application: Macroeconomic Models

One particularly well-known example of these ideas in action is the [Lucas Critique](#), a famous 1976 paper by [Bob Lucas](#) critiquing the use of simple statistical models for evaluation of economic policy decisions. Lucas' paper gives several broad examples, but arguably the most remembered example is policy decisions based on the [Phillips curve](#).

The Phillips curve is an empirical relationship between unemployment and inflation. Phillips examined almost a century of economic data, and showed a consistent negative correlation: when inflation was high, unemployment was low, and vice-versa. In other words, prices and wages rise faster at the peak of the business cycle (when unemployment is low) than at the trough (when unemployment is high).

The obvious mistake one might make, based on the Phillips curve, is to think that perpetual low unemployment can be achieved simply by creating perpetual inflation (e.g. by printing money). Lucas opens his critique by eviscerating this very idea:

The inference that permanent inflation will therefore induce a permanent economic high is no doubt [...] ancient, yet it is only recently that this notion has undergone the mysterious transformation from obvious fallacy to cornerstone of the theory of economic policy.

Bear in mind that this was written in the mid-1970's - the era of "stagflation", when both inflation and unemployment were high for several years. Stagflation was an empirical violation of the Phillips curve - the historical behavior of the system broke down when central banks changed their policies to pursue more inflation, and people changed their behavior to account for faster expected inflation in the future.

In short: a statistical model with no gears in it completely fell apart when one part of the system (the central bank) changed its behavior.

On the other hand, *before* stagflation was under way, multiple theorists (notably [Edmund Phelps](#) and [Milton Friedman](#), via very different approaches) published simple gears-level models of the Phillips curve which predicted that it would break down if currencies were predictably devalued - i.e. if people expected central banks to print more money. The key "gears" in these models were individual agents - the macroeconomic behavior (unemployment-inflation relationship) was explained in terms of the expectations and decisions of all the individual people.

This led to a paradigm shift in macroeconomics, beginning the era of "microfoundations": macroeconomic models derivable from microeconomic models of the expectations and behavior of individual agents - in other words, gears-level models of the economy.

Gears from Behavior?

In general, we cannot *fully* learn gears-level models by looking only at externally-visible input-output behavior. Our hypothetical 3- or 5-gear boxes are a case in point.

However, some kinds of models can at least deduce *something* about gears-level structure by looking at externally-visible behavior.

For example: given a gearbox with a crank and wheel, it's entirely possible that the rotation of the wheel has hysteresis, a.k.a. memory - it depends not only on the crank's rotation now,

but also the crank's rotation earlier. This would be the case if, for instance, the box contains a flywheel. If we look at the data and see that the wheel's rotation has no dependence on the crank's rotation at earlier times (after accounting for the crank's current rotation), then we can conclude that the box probably does not contain any flywheels or other hysteretic components (or if it does, they're small or decoupled from the wheel).

More generally, these sort of conditional independence relationships fall under the umbrella of [probabilistic causal models](#). By testing different causal models on externally-visible data, we can back out information about the internal cause-and-effect structure of the system. If we see that only the crank's *current* rotation matters to the wheel, then that rules out internal components with memory.

Causal models are the largest class of statistical models I know of which yield information about internal gears. However, they're not the only way to build gears-level models from behavior. If we have strong prior information, we can often use behavioral data to directly compare gears-level hypotheses.

Application: Wolf's Dice

Around the mid-19th century, Swiss astronomer Rudolf Wolf rolled a pair of dice 20000 times, recording each outcome. The main result was that the dice were definitely not perfectly fair - there were small but statistically significant biases.

Now, we could easily look at Wolf's data and use it to estimate the frequency with which each face of each die is rolled. But that's not a gears-level model; it doesn't say anything about the physical die.

In order to back out gears-level information from the data, we need to leverage our prior knowledge about dice and die-making. Jaynes did exactly this in [a 1979 paper](#); the key pieces of prior information are:

- We know dice are roughly cube-shaped, and any difference in face frequencies should stem from asymmetry of the physical die. We know 3 is opposite 4, 2 is opposite 5, and 1 is opposite 6.
- We know dice have little pips showing the numbers on each face; different faces have different numbers of pips, which we'd expect to introduce a slight asymmetry.
- Imagining how the dice might have been manufactured, Jaynes guesses that the final cut would have been more difficult to make perfectly even than the earlier cuts - leaving one axis slightly shorter/longer than the other two.

Based on those asymmetries, we'd guess:

- One of the three face pairs (3, 4), (2, 5), (1, 6) has significantly different frequency from the others, corresponding to the last axis cut.
- The faces with fewer pips (3, 2, and especially 1) have slightly lower frequency than those with more pips (4, 5, and especially 6), since more pips means slightly less mass near that face.
- Other than that, the frequencies should be pretty even.

This is basically a guess-and-check process: we guess what asymmetry might be present based on our prior knowledge, consider how that would change the behavior, then we use the data to check the model.

Jaynes tests out these models, and finds that (1) the white die's 3-4 axis is slightly shorter than the other two, and (2) the pips indeed shift the center of mass slightly away from the center of the die. These two asymmetries together explain all of the bias seen in the data, so the die should be quite symmetric otherwise. I analyze the same problem in [this post](#) (using slightly different methods from Jaynes) and reproduce the same result.

Because this is a gears-level model, we could in principle check the result using a “side channel”: if we could track down the dice Wolf used, then we could take out our calipers and measure the lengths of the 3-4, 2-5, and 1-6 axes. Our prediction is that the 2-5 and 1-6 axes would be close, but the 3-4 axis would be significantly shorter. Note that we still don’t have a *full* gears-level model - we don’t predict *how much* shorter the 3-4 axis is. We don’t have a way to back out all the dimensions of the die. But we certainly expect the difference between the 3-4 length and the 2-5 length to be much larger than the difference between the 2-5 length and the 1-6 length. Our model yields *some* information about gears-level structure.

Takeaway

Statistics, machine learning, and adjacent fields tend to have a myopic focus on predicting future data.

Gears-level models cannot be fully learned by looking at externally-visible behavior data. That makes it hard to prove theorems about convergence of statistical methods, or write tests for machine learning algorithms, when the goal is to learn about a system’s internal gears. So, to a large extent, these fields have ignored gears-level learning and focused on predicting future data. Gears have snuck in only to the extent that they’re useful for predicting externally-visible behavior.

But sooner or later, any field dominated by a gears-less worldview will have its Lucas Critique.

It is possible to leverage probability to test gears-level models, and to back out at least some information about a system’s internal structure. It’s not easy. We need to restrict ourselves to certain classes of models (i.e. causal models) and/or leverage lots of prior knowledge (e.g. about dice). It looks less like black-box statistical/ML models, and more like science: think about what the physical system might look like, figure out how the data would differ between different possible physical systems, and then go test it. The main goal is not to predict future data, but to compare models.

That’s the kind of approach we need to build models which won’t fall apart every time central banks change their policies.

Long-term Donation Bunching?

Let's say you're a relatively well-off American with an income of \$100k and have taken the [Giving What We Can](#) pledge to donate 10% of your money to the most effective charities you can find. The standard deduction in the US is currently \$12k, which means if you were to donate \$10k/year it effectively wouldn't be tax deductible. [1]

The standard EA suggestion here is [Donation Bunching](#): instead of donating \$10k every year donate \$20k half the years and \$0 the other half. In the years where you donate you itemize your deductions, in the others you take the standard deduction. Considered over two years, \$8k of the \$20k (40%) is deductible.

But we could do better than that! You can deduct up to 60% of your income if you're donating cash, so you could have five years of donating \$0 and one year of donating \$60k. Considered over six years, \$48k of the \$60k (80%) is deductible. [2]

This sounds great, right? Money donated to effective charities does much more good than money collected by the government, and with a bit of planning you can effectively move 80% of your donations to being pre-tax. But I don't think it's a good idea!

Over time people change, and a very common way people change as they get older is to turn inward. You start out as a bright-eyed idealist, enthusiastic about making the world a better place and willing to make sacrifices for what you believe in. Then you burn out working too many hours doing something that doesn't feel as effective as you thought it would be, or you start feeling a pull to have kids and focus your efforts there, or you just stop feeling so motivated by altruism, or dozens of other things, and after a few years the idea of giving your money to people who need it more still sounds nice but isn't a priority anymore.

While we don't have good data on the rate at which this happens, in a [small sample](#) about half of people in the effective altruism movement in 2013 were no longer involved five years later. If you think your current self is correct to be altruistic and don't want to leave donations up to a likely less-generous future person then bunching donations over several years is harmful: the substantial possibility that you don't actually donate outweighs the tax savings. [3]

(Thanks to someone I talked to at the [SSC Boston Meetup](#) for asking me the question that got me thinking about this.)

[1] Specifically, unless you had other reasons to itemize your tax deductions, you would do better to take the \$12k standard deduction.

[2] This ignores inflation and investment income, but they aren't large enough to change the picture much over a ~6 year window.

[3] Depending on your views on discount rates ("how much better is it to give this year than next?") it might also not be so good.

Comment via: [facebook](#)

How good is the case for retraining yourself to sleep on your back?

I frequently hear the advice that it's better to sleep on the back and worthwhile to learn to sleep on your back. Are there any studies that backup that advice. Otherwise are there other good arguments? Personal experience is also welcome.

Specificity: Your Brain's Superpower



Introduction

What if there was a magic superpower that simultaneously enabled you to:

- [Demolish bad arguments](#)
- [Judge startup ideas](#)
- [Make scientific breakthroughs](#)
- [Solve climate change](#)
- [Understand "God"](#)
- [Be emotionally mature](#)
- [Teach concepts better](#)
- [Draw better](#)
- Be creative
- Make small talk
- Improve discourse
- Be invisible

Amazingly, your brain *does* have a superpower that gives you all these powers. It's called...

SPECIFICITY

To activate the power of specificity, all you have to do is ask yourself the question, "**What's an example of that?**" Or more bluntly, "**Can I be more specific?**" And

then you unleash a ton of power in a surprisingly broad variety of domains.

It's an open secret in the rationality community how powerful this skill is of being specific. Eliezer captured the essence of it in his 2012 post, [Be Specific](#). In it, he comments on the difficulty of teaching people specificity skills for the first time:

When I'm talking to anyone outside the local LessWrong community, I find that a very large amount of my conversation involves repeatedly asking them to be more specific.

He also describes how CFAR's [Applied Rationality Workshops](#) teach the power of specificity by osmosis. I can vouch that my personal experience as a workshop attendee fits this description:

Attendees picked [specificity] up from all the instructors having to repeatedly ask the attendees to be more specific, and then having to ask them again, while being specific themselves, until the attendees picked up the rhythm by example and feedback.

I hope you're curious to unpack my list of claims about what specificity lets you do, because I'm *dead serious* about all of it! Except being invisible; I was kidding about that one.

So without further ado, let's get into the specifics.

Next post: [The Power to Demolish Bad Arguments](#)

Idols of the Mind Pt. 1 (Novum Organum Book 1: 38-52)

This is the fourth post in the [Novum Organum sequence](#). For context, see [the sequence introduction](#).

We have used Francis Bacon's *Novum Organum* in the version presented at www.earlymoderntexts.com. Translated by and copyright to [Jonathan Bennett](#). Prepared for LessWrong by [Ruby](#).

Ruby's Reading Guide

Novum Organum is organized as two books each containing numbered "aphorisms." These vary in length from three lines to sixteen pages. Titles of posts in this sequence, e.g. *Idols of the Mind Pt. 1*, are my own and do not appear in the original.

While the translator, Bennett, encloses his editorial remarks in a single pair of [brackets], I have enclosed mine in a [[double pair of brackets]].

Bennett's Reading Guide

[Brackets] enclose editorial explanations. Small ·dots· enclose material that has been added, but can be read as though it were part of the original text. Occasional •bullets, and also indenting of passages that are not quotations, are meant as aids to grasping the structure of a sentence or a thought. Every four-point ellipsis indicates the omission of a brief passage that seems to present more difficulty than it is worth. Longer omissions are reported between brackets in normal-sized type.

Aphorism Concerning the Interpretation of Nature: Book 1: 38-52

by Francis Bacon

38. The idols and false notions that now possess the human intellect and have taken deep root in it don't just •occupy men's minds so that truth can hardly get in, but also when a truth *is* allowed in they will •push back against it, stopping it from contributing to a fresh start in the sciences. This can be avoided only if men are forewarned of the danger and do what they can to fortify themselves against the assaults of these idols and false notions.

39. There are four classes of idols that beset men's minds, and to help me in my exposition I have given them names. I call the first class **idols of the tribe**, the second **idols of the cave**, the third **idols of the market place**, and the fourth **idols of the theatre**.

40. The proper way to keep idols at bay and to drive them off is, no doubt, to form ideas and axioms by true induction. But it is very useful just to point the idols out; for •the truth about the idols serves •the interpretation of nature in the way that •the truth about argumentative fallacies serves •ordinary logical argumentation.

41. The **idols of the tribe** have their foundation in human nature itself—in the tribe known as 'mankind'. It is not true that the human senses are the measure of things; for all perceptions—of the senses as well as of the mind—reflect the perceiver rather than the world. The human intellect is like a distorting mirror, which receives light-rays irregularly and so mixes its own nature with the nature of things, which it distorts.

[[This is something of a reference to the [Mind Projection Fallacy](#).]]

42. The **idols of the cave** are the idols of the individual man. In addition to the errors that are common to human nature in general, everyone has his own personal cave or den that breaks up and corrupts the light of nature. This may come from factors such as these:

- his own individual nature,
- how he has been brought up and how he interacts with others,
- his reading of books and the influence of writers he esteems and admires,
- differences in how his environment affects him because of differences in his state of mind—whether it is busy thinking about something else and prejudiced against this intake or calm and open-minded.

So that the human spirit is distributed among individuals in ways that make it variable and completely disorderly—almost a matter of luck. Heraclitus was right: men look for sciences in ·their own individual· lesser worlds, and not in the greater world that they have in common.

[[Related: [Epistemic Luck](#). However, I believe Bacon's eventual thesis is that even though luck may determine your starting point, proper use of tools, i.e. empiricism,

can lead to correct conclusions even if you started out unlucky.]]

43. There are also idols formed by men's agreements and associations with each other (·I have in mind especially the agreements that fix the meanings of words·). I call these **idols of the market place**, because that is where men come together and do business. ·Such transactions create idols· because •men associate by talking to one another, and •the uses of words reflect common folks' ways of thinking. It's amazing how much the intellect is hindered by wrong or poor choices of words. The definitions or explanations that learned men sometimes use to protect themselves ·against such troubles· don't at all set the matter right: words plainly force and overrule the intellect, throw everything into confusion, and lead men astray into countless empty disputes and idle fancies.

[[Bacon grokked that misuses of words were a great cause of confusion. He probably would have like the [A Human's Guide to Words Sequence](#). See [Where to Draw the Boundary?](#) and [37 Ways That Words Can Be Wrong.](#)]]

44. Lastly, there are idols that have come into men's minds from various philosophical dogmas and from topsy-turvy laws of demonstration. I call these **idols of the theatre**, because I regard every one of the accepted systems as the staging and acting out of a fable, making a fictitious staged world of its own. I don't say this only about the systems that are currently fashionable, or only about the ancient sects and philosophies; many other fables of the same kind may still be written and produced, seeing that errors can be widely different yet have very similar causes. And I'm saying this not only about whole systems but also about a good many principles and axioms in ·individual· sciences—ones that have gathered strength through tradition, credulity, and negligence. But these various kinds of idols will have to be discussed more clearly and at greater length if the human intellect is to be adequately warned against them. ·I'll start with the idols of the tribe, which will be my topic until the end of **52**·.

45. The human intellect is inherently apt to •suppose the existence of more order and regularity in the world than it •finds there. Many things in nature are unique and not like anything else; but the intellect devises for them non-existent parallels and correspondences and relatives. That is how it comes about •that all the heavenly bodies are thought to move in perfect circles. . . . , •that fire. . . has been brought in as one of the elements, to complete the square with the other three elements—·earth, air, water·—which the senses detect, and •that the 'elements' (as they are called) are arbitrarily said to differ in density by a factor of ten to one. And so on for other dreams. And these fancies affect not only ·complex· propositions but also simple notions.

[[[People see patterns everywhere](#), many that aren't there.]]

46. Once a human intellect has adopted an opinion (either as something it *likes* or as something generally accepted), it draws everything else in to confirm and support it. [[Confirmation bias.]] Even if there are more and stronger instances against it ·than there are in its favour·, the intellect either •overlooks these or •treats them as negligible or •does some line-drawing that lets it shift them out of the way and reject them. This involves a great and pernicious prejudgment by means of which the intellect's former conclusions remain inviolate.

A man was shown a picture, hanging in a temple, of people who had made their vows and escaped shipwreck, and was asked 'Now do you admit the power of the

gods?' He answered with a question: 'Where are the pictures of those who made their vows and then drowned?'

[[A correct identification of [selection bias/survivorship bias/anthropic bias.](#)]]

It was a good answer! That's how it is with all superstition— involving astrology, dreams, omens, divine judgments, and the like, Men get so much pleasure out of such vanities that they notice the •confirming events and inattentively pass by the more numerous •disconfirming ones. This mischief insinuates itself more subtly into philosophy and the sciences: there, when a proposition has found favour it colours other propositions and brings them into line with itself, even when they ·in their undisguised form· are sounder and better than it is. Also, apart from the pleasure and vanity that I have spoken of, the human intellect is perpetually subject to the special error of being moved and excited more by affirmatives than by negatives; whereas it *ought* to have the same attitude towards each. Indeed, when it is a matter of establishing a true axiom, it's the negative instance that carries more force.

[[The idea of looking for disconfirming negative instances is expounded in [Positive Bias: Look Into the Dark.](#).]]

47. The greatest effect on •the human intellect is had by things that strike and enter the mind simultaneously and unexpectedly; it is these that customarily fill—*inflate!*—the imagination; and then •it feigns and supposes that everything else is somehow, though •it can't see how, similar to those few things that have taken it by storm. ['Feign' translates the Latin *fingo*, which is the source for the English word 'fiction'.] But the intellect is altogether slow and unfit for the journey to distant and heterogeneous instances which put axioms to the test—like testing something by fire —unless it is forced to do so by severe laws and overruling authority.

48. The human intellect is never satisfied; it can't stop or rest, and keeps searching further; but all to no purpose. That's why we can't conceive of any end or limit to the world—why we always virtually *have* to have the thought of something beyond ·any candidate for the role of world's end·. And we can't conceive, either, of how eternity has flowed down to the present day. ·A plausible story about this says that time is infinite in both directions, and the present is just a point along this infinite line. But the commonly accepted idea of infinity in time past and in time to come can't be sustained, for it implies that •one infinity is greater than another, and that •one infinity is getting used up and tending to become finite. The infinite divisibility of lines is a source of a similar network of difficulties arising from our thought's inability ·to reach a resting-place·. But this inability interferes even worse in the discovery of causes, ·and here is how·.

The most general principles in nature have to be brute facts, just as they are discovered, and can't be derived from any ·still more general or basic· cause. Yet the restless human intellect still looks for something

(**Latin:** *notiora* = 'better known', **probably short for:** *natura notiora* = 'better known to nature', **actually meaning:** 'more general and/or basic' [see note in 22])— ·something to explain why they are true·.

Then in that ·doomed· struggle for something further off, it ·finds itself defeated, and instead· falls back on something that is nearer at hand, namely on *final causes*—·i.e. on the notion of what a principle is *for*, what *purpose* explains its being true·. Science has been enormously messed up by this appeal to final causes, which obviously come

from the nature of man rather than from the nature of the world—that is, which project the scientist's own purposes *onto* the world rather than finding purposes *in it*.

[[Similarly another case of [Mind Projection Fallacy](#).]]

To look for causes of the most general principles is to do science in an ignorant and frivolous way—just as much as *not* looking for causes of subordinate and less general truths.

49. The human intellect doesn't burn with a dry [here = 'uncontaminated'] light, because what the person *wants* and *feels* gets pumped into it; and that is what gives rise to the 'please-yourself sciences'. For a man is more likely to believe something if he would like it to be true. Therefore he rejects:

- difficult things because he hasn't the patience to research them,
- sober and prudent things because they narrow hope,
- the deeper things of nature, from superstition,
- the light that experiments can cast, from arrogance and pride (not wanting people to think his mind was occupied with trivial things),
- surprising truths, out of deference to the opinion of the vulgar.

In short, there are countless ways in which, sometimes imperceptibly, a person's likings colour and infect his •intellect.

[[This aphorism calls out the general behavior of [motivated cognition](#).]]

50. But what contributes most to the blockages and aberrations of the human intellect is the fact that the •human• senses are dull, incompetent and deceptive. The trouble is this: things that strike the senses outweigh other things—more important ones—that don't immediately strike them. That is why people stop *thinking* at the point where their *eyesight* gives out, paying little or no attention to •things that can't be seen—for example, all the •workings of the spirits enclosed in tangible bodies. Nor do they pay attention to all the subtler changes of microstructure in the parts of coarser substances (which are vulgarly called 'alterations' though they are really extremely small-scale •movements). And yet unless these two things—the workings of spirits, and subtle changes of form in bodies—can be searched out and brought into the light, nothing great can be achieved in nature in the way of practical applications. A third example: the essential nature of our common air, and of all the many bodies that are less dense than air, is almost unknown. For the senses by themselves are weak and unreliable; and instruments for extending or sharpening them don't help much. All the truer kind of *interpretation* of nature comes about through instances and well-designed experiments: the senses pass judgment on the experiment, and the experiment passes judgment on nature, on the facts.

[Bacon's many uses of the word *schematismus* show that for him a body's *schematismus* is its fine-grained structure. This version will always use 'microstructure', but be aware that Bacon doesn't use a word with the prefix 'micro'. •Also, here and throughout, 'spirits' are extremely finely divided gases or fluids, *not* mental items of any kind.]

51. The human intellect is inherently prone to make abstractions, and it feigns an unchanging essence for things that are in flux. But better than •abstracting from nature is •dissecting it; which is what Democritus and his followers did, getting deeper into nature than anyone since. What we should be attending to is *matter*, its microstructures and changes of microstructure, and *actus purus*, and the laws of

action or motion. ·The alternative to studying *matter* is to study *forms*, but· forms are fabrications of the human mind, unless you want to call the laws of action ‘forms’.

[Bacon doesn’t explain *actus purus*. In each of its other three occurrences he connects it with *laws*, and his meaning seems to be something like: ‘the laws governing the pure actions of individual things, i.e. the things they do because of their own natures independently of interference from anything else’. If x does A partly because of influence from something else y, then x is not purely •active in respect of A because y’s influence gives A a certain degree of •passivity. From here on, *actus purus* will be translated by ‘pure action’.]

52. Those, then, are the idols of the tribe, as I call them— the idols that ·arise from human nature as such. More specifically, they· arise from the human spirit’s •regularity of operation, or its •prejudices, or its •narrowness, or its •restlessness, or •input from the feelings, or from the •incompetence of the senses, or from •the way the senses are affected.

The next post in the sequence, Book 1: 53-68 (Idols of the Mind Pt. 2), will be posted Thursday, September 26 at latest by 4:00pm PDT.

Realism and Rationality

Format warning: This post has somehow ended up consisting primarily of substantive endnotes. It should be fine to read just the (short) main body without looking at any of the endnotes, though. The endnotes elaborate on various claims and distinctions and also include a much longer discussion of decision theory.

Thank you to Pablo Stafforini, Phil Trammell, Johannes Treutlein, and Max Daniel for comments on an initial draft. I have also slightly edited the post since I first published it, to try to make a few points clearer.

When discussing normative questions, it is not uncommon for members of the rationalist community to identify as anti-realists. But normative anti-realism seems to me to be in tension with some of the community's core interests, positions, and research activities. In this post I suggest that the cost of rejecting realism may be larger than is sometimes recognized. [\[1\]](#)

1. Realism and Anti-Realism

Everyone is, at least sometimes, inclined to ask: "What should I do?"

We ask this question when we're making a decision and it seems like there are different considerations to be weighed up. You might be considering taking a new job in a new city, for example, and find yourself wondering how to balance your preferences with those of your significant other. You might also find yourself thinking about whether you have any obligation to do impactful work, about whether it's better to play it safe or take risks, about whether it's better to be happy in the moment or to be able to look back with satisfaction, and so on. It's almost inevitable that in a situation like this you will find yourself asking "What should I do?" and reasoning about it as though the question has an answer you can approach through a certain kind of directed thought. [\[2\]](#)

But it's also conceivable that this sort of question doesn't actually have an answer. Very roughly, at least to certain philosophers, *realism* is a name for the view that there are some things that we should do or think. *Anti-realism* is a name for the view that there are not. [\[3\]](#)[\[4\]](#)[\[5\]](#)[\[6\]](#)

2. Anti-Realism and the Rationality Community

In discussions of normative issues, it seems not uncommon for members of the rationalist community to identify as "anti-realists." Since people in different communities can obviously use the same words to mean different things, I don't know what fraction of rationalists have the same thing in mind when they use the term "anti-realism."

To the extent people do have the same thing in mind, though, I find anti-realism hard to square with a lot of other views and lines of research that are popular within the community. A few main points of tension stand out to me.

2.1 Normative Uncertainty

One first point of tension is the community's relatively strong interest in the subject of normative uncertainty. At least as it's normally discussed in the philosophy literature,

normative uncertainty is uncertainty about normative facts that bear on what we should do. If we assume that anti-realism is true, though, then we are assuming that there are no such facts. It seems to me like a committed anti-realist could not be in a state of normative uncertainty.

It may still be the case, as [Sepielli \(2012\)](#) suggests, that a committed anti-realist can experience psychological states that are interestingly *structurally analogous* to states of normative uncertainty. However, [Bykvist and Olson \(2012\)](#) disagree (in my view) fairly forcefully, and Sepielli is in any case clear that: "Strictly speaking, there cannot be such a thing as normative uncertainty if non-cognitivism [the dominant form of anti-realism] is true."^[7]

2.2 Strongly Endorsed Normative Views

A second point of tension is the existence of a key set of normative claims that a large portion of the community seems to treat as true.

One of these normative claims is the Bayesian claim that we ought to have degrees of belief in propositions that are consistent with the Kolmogorov probability axioms and that are updated in accordance with Bayes' rule. It seems to me like very large portions of the community self-identify as Bayesians and regard other ways of assigning and updating degrees of belief in propositions as not just different but incorrect.

Another of these normative claim is the subjectivist claim that we should do whatever would best fulfill some version of our current preferences. To learn what we should do, on this view, the main thing is to introspect about our own preferences.^[8] Whether or not a given person should commit a violent crime, for instance, depends purely on whether they want to commit the crime (or perhaps on whether they would want to commit it if they went through some particular process of reflection).

A further elaboration on this claim is that, when we are uncertain about the outcomes of our actions, we should more specifically act to maximize the *expected* fulfillment of our desires. We should consider the different possible outcomes of each action, assign them probabilities, assign them desirability ratings, and then use the expected value formula to rate the overall goodness of the action. Whichever action has the best overall rating is the one we should take.

One possible way of squaring an endorsement of anti-realism with an apparent endorsement of these normative claims is to argue that people don't actually have normative claims in mind when they write and talk about these issues. Non-cognitivists -- a particular variety of anti-realists -- argue that many utterances that seem at first glance like claims about normative facts are in fact nothing more than expressions of attitudes. For instance, an [emotivist](#) -- a further sub-variety of non-cognitivist -- might argue that the sentence "You should maximize the expected fulfillment of your current desires!" is simply a way of expressing a sense of fondness toward this course of action. The sentence might be cached out as being essentially equivalent in content to the sentence, "Hurrah, maximizing the expected fulfillment of your current desires!"

Although a sizeable portion of philosophers are non-cognitivists, I generally don't find it very plausible as a theory of what people are trying to do when they seem to make normative claims.^[9] In this case it doesn't feel to me like most members of the rationalist community are just trying to describe one particular way of thinking and

acting, which they happen to prefer to others. It seems to me, rather, that people often talk about updating your credences in accordance with Bayes' rule and maximizing the expected fulfillment of your current desires as the *correct* things to do.

One more thing that stands out to me is that arguments for anti-realism often seem to be presented as though they implied (rather than negated) the truth of some of these normative claims. For example, the popular "Replacing Guilt" sequence on Minding Our Way seems to me to repeatedly attack normative realism. It [rejects](#) the idea of "shoulds" and [points out](#) that there aren't "any oughtthorities to ordain what is right and what is wrong." But then it seems to draw normative implications out of these attacks: among other implications, you should "just do what you want." At least taken at face value, this line of reasoning wouldn't be valid. It makes no more sense than reasoning that, if there are no facts about what we should do, then we should "just maximize total hedonistic well-being" or "just do the opposite of what we want" or "just open up souvenir shops." Of course, though, there's a good chance that I'm misunderstanding something here.

2.3 Decision Theory Research

A third point of tension is the community's engagement with normative decision theory research. Different normative decision theories pick out different necessary conditions for an action to be the one that a given person should take, with a focus on how one should respond to uncertainty (rather than on what ends one should pursue). [\[10\]](#)[\[11\]](#)

A typical version of [CDT](#) says that the action you should take at a particular point in time is the one that would *cause* the largest expected increase in value (under some particular framework for evaluating causation). A typical version of [EDT](#) says that the action you should take at a particular point in time is the one that would, once you take it, allow you to rationally expect the most value. There are also alternative versions of these theories -- for instance, versions using [risk-weighted expected value maximization](#) or the criterion of [stochastic dominance](#) -- that break from the use of pure expected value.

I've pretty frequently seen it argued within the community (e.g. in the papers "[Cheating Death in Damascus](#)" and "[Functional Decision Theory](#)") that CDT and EDT are not "correct" and that some other new theory such as [functional decision theory](#) is. But if anti-realism is true, then no decision theory is correct.

Eliezer Yudkowsky's influential [early writing](#) on decision theory seems to me to take an anti-realist stance. It suggests that we can only ask meaningful questions about the effects and correlates of decisions. For example, in the context of the Newcomb thought experiment, we can ask whether one-boxing is correlated with winning more money. But, it suggests, we cannot take a step further and ask what these effects and correlations imply about what it is "reasonable" for an agent to do (i.e. what they *should* do). This question -- the one that normative decision theory research, as I understand it, is generally about -- is seemingly dismissed as vacuous.

If this apparently anti-realist stance is widely held, then I don't understand why the community engages so heavily with normative decision theory research or why it takes part in discussions about which decision theory is "correct." It strikes me a bit like an atheist enthusiastically following theological debates about which god is the true god. But I'm mostly just confused here. [\[12\]](#)[\[13\]](#)

3. Sympathy for Realism

I wouldn't necessarily describe myself as a realist. I get that realism is a weird position. It's both metaphysically and epistemologically suspicious. What is this mysterious property of "should-ness" that certain actions are meant to possess -- and why would our intuitions about which actions possess it be reliable?[\[14\]](#)[\[15\]](#)

But I am also very sympathetic to realism and, in practice, tend to reason about normative questions as though I was a full-throated realist. My sympathy for realism and tendency to think as a realist largely stems from my perception that if we reject realism and internalize this rejection then there's really not much to be said or thought about anything. We can still express attitudes at one another, for example suggesting that we like certain actions or credences in propositions better than others. We can present claims about the world, without any associated explicit or implicit belief that others should agree with them or respond to them in any particular way. And that seems to be about it.

Furthermore, if anti-realism is true, then it can't also be true that we should believe that anti-realism is true. Belief in anti-realism seems to undermine itself. Perhaps belief in realism is self-undermining in a similar way -- if seemingly correct reasoning leads us to account for all the ways in which realism is a suspect position -- but the negative feedback loop in this case at least seems to me to be less strong.[\[16\]](#)

I think that realism warrants more respect than it has historically received in the rationality community, at least relative to the level of respect it gets from philosophers.[\[17\]](#) I suspect that some of this lack of respect might come from a relatively weaker awareness of the cost of rejecting realism or of the way in which belief in anti-realism appears to undermine itself.

-
1. I'm basing my views I express in this post primarily off Derek Parfit's writing, specifically his book *On What Matters*. For this reason, it seems pretty plausible to me that there are some important points I've missed by reading too narrowly. In addition, it also seems likely that some of the ways in which I talk about particular issues around normativity will sound a bit foreign or just generally "off" to people who are highly familiar with some of these issues. One unfortunate reason for this is that the study of normative questions and of the nature of normativity seems to me to be spread out pretty awkwardly across the field of philosophy, with philosophers in different sub-disciplines often discussing apparently interconnected questions in significant isolation of one another while using fairly different terminology. This means that (e.g.) meta-ethics and decision theory are seldom talked about at the same time and are often talked about in ways that make it difficult to see how they fit together. A major reason I am leaning on Parfit's work is that he is -- to my knowledge -- one of relatively few philosophers to have tried to approach questions around normativity through a single unified framework. [←](#)
 2. This is a point that is also discussed at length in David Enoch's book *Taking Morality Seriously* (pgs. 70-73):

Perhaps...we are essentially deliberative creatures. Perhaps, in other words, we cannot avoid asking ourselves what to do, what to believe, how to reason, what to care about. We can, of course, stop deliberating about one thing or another, and it's not as if all of us have to be practical philosophers

(well, if you're reading this book, you probably are, but you know what I mean). It's opting out of the deliberative project as a whole that may not be an option for us....

[Suppose] law school turned out not to be all you thought it would be, and you no longer find the prospects of a career in law as exciting as you once did. For some reason you don't seem to be able to shake off that old romantic dream of studying philosophy. It seems now is the time to make a decision. And so, alone, or in the company of some others you find helpful in such circumstances, you deliberate. You try to decide whether to join a law firm, apply to graduate school in philosophy, or perhaps do neither.

The decision is of some consequence, and so you resolve to put some thought into it. You ask yourself such questions as: Will I be happy practicing law? Will I be happier doing philosophy? What are my chances of becoming a good lawyer? A good philosopher? How much money does a reasonably successful lawyer make, and how much less does a reasonably successful philosopher make? Am I, so to speak, more of a philosopher or more of a lawyer? As a lawyer, will I be able to make a significant political difference? How important is the political difference I can reasonably expect to make? How important is it to try and make any political difference? Should I give any weight to my father's expectations, and to the disappointment he will feel if I fail to become a lawyer? How strongly do I really want to do philosophy? And so on. Even with answers to most – even all – of these questions, there remains the ultimate question. "All things considered", you ask yourself, "what makes best sense for me to do? When all is said and done, what should I do? What shall I do?"

When engaging in this deliberation, when asking yourself these questions, you assume, so it seems to me, that they have answers. These answers may be very vague, allow for some indeterminacy, and so on. But at the very least you assume that some possible answers to these questions are better than others. You try to find out what the (better) answers to these questions are, and how they interact so as to answer the arch-question, the one about what it makes most sense for you to do. You are not trying to create these answers. Of course, in an obvious sense what you will end up doing is up to you (or so, at least, both you and I are supposing here). And in another, less obvious sense, perhaps the answer to some of these questions is also up to you. Perhaps, for instance, how happy practicing law will make you is at least partly up to you. But, when trying to make up your mind, it doesn't feel like just trying to make an arbitrary choice. This is just not what it is like to deliberate. Rather, it feels like trying to make the right choice. It feels like trying to find the best solution, or at least a good solution, or at the very least one of the better solutions, to a problem you're presented with. What you're trying to do, it seems to me, is to make the decision it makes most sense for you to make. Making the decision is up to you. But which decision is the one it makes most sense for you to make is not. This is something you are trying to discover, not create. Or so, at the very least, it feels like when deliberating.



3. Specifically, the two relevant views can be described as realism and anti-realism with regard to "normativity." We can divide the domain of "normativity" up into

the domains of “[practical rationality](#),” which describes what actions people should take, and “epistemic rationality,” which describes which beliefs or degrees of belief people should hold. The study of ethics, decision-making under uncertainty, and so on can then all be understood as sub-components of the study of practical rationality. For example, one view on the study of ethics is that it is the study of how factors other than one’s own preferences might play roles in determining what actions one should take. It should be noted that terminology varies very widely though. For example, different authors seem to use the word “ethics” more or less inclusively. The term “moral realism” also sometimes means roughly the same thing as “normative realism,” as I’ve defined it here, and sometimes picks out a more specific position. ↪

4. An edit to the initial post, I think it's probably worth saying more about the concept of "moral realism" in relation to "normative realism." Depending on the context, "moral realism" might be taken to refer to: (a) normative realism, (b) realism about practical rationality (not just epistemic rationality), (c) realism about practical rationality combined with the object-level belief that people should do more than just try to satisfy their own personal preferences, or (d) something else in this direction.

One possible reason the term lacks a consensus definition is that, perhaps surprisingly, many contemporary “moral realists” aren’t actually very preoccupied with the concept of “morality.” Popular books like *Taking Morality Seriously*, *On What Matters*, and *The Normative Web* spend most of their energy defending normative realism, more broadly, and my impression is that their critics spend most of their energy attacking normative realism more broadly. One reason for this shift in focus toward normative realism is the realization that, on almost any conception of “moral realism,” nearly all of the standard metaphysical and epistemological objections to “moral realism” also apply just as well to normative realism in general. Another reason is that any possible distinction between moral and normative-but-not-moral facts doesn’t seem like it could have much practical relevance: If we know that we should make some decision, then we know that we should take it; we have no obvious additional need to know or care whether this normative fact warrants the label “moral fact” or not. Here, for example, is David Enoch, in *Taking Morality Seriously*, on the concept of morality (pg. 86):

What more...does it take for a normative truth (or falsehood) to qualify as moral? Morality is a particular instance of normativity, and so we are now in effect asking about its distinctive characteristics, the ones that serve to distinguish between the moral and the rest of the normative. I do not have a view on these special characteristics of the moral. In fact, I think that for most purposes this is not a line worth worrying about. The distinction within the normative between the moral and the non-moral seems to me to be shallow compared to the distinction between the normative and the non-normative - both philosophically, and, as I am about to argue, practically. (Once you know you have a reason to X and what this reason is, does it really matter for your deliberation whether it qualifies as a *moral reason*?)

↪

5. There are two major strands of anti-realism. Error theory (sometimes equated with “nihilism”) asserts that all claims that people should do particular things or refrain from doing particular things are false. Non-cognitivism asserts that

utterances of the form “A should do X” typically cannot even really be understood as claims; they’re not the sort of thing that could be true or false. ↵

6. In this post, for simplicity, I’m talking about normativity using binary language. Either it’s the case that you “should” take an action or it’s not the case that you “should” take it. But we might also talk in less binary terms. For example, there may be some actions that you merely have “more reason” to take than others. ↵
7. In Sepielli’s account, for example, the experience of feeling extremely in favor of blaming someone a little bit for taking an action X is analogous to the experience of being extremely confident that it is a little bit wrong to take action X. This account is open to at least a few objections, such as the objection that degrees of favorability don’t -- at least at first glance -- seem to obey the standard axioms of probability theory. Even if we do accept the account, though, I still feel unclear about the proper method and justification for converting debates around normative uncertainty into debates around these other kinds of psychological states. ↵
8. If my memory is correct, one example of a context in which I have encountered this subjectivist viewpoint is in a CFAR workshop. One lesson instructs attendees that if it seems like they “should” do something, but then upon reflection they realize they don’t want to do it, then it’s not actually true that they should do it. ↵
9. The PhilPapers survey suggests that about a quarter of both normative ethicists and applied ethicists also self-identify as anti-realists, with the majority of them presumably leaning toward non-cognitivism over error theory. It’s still an active [matter of debate](#) whether non-cognitivists have sensible stories about what people are trying to do when they seem to be discussing normative claims. For example, naive emotivist theories stumble in trying to explain sentences like: “It’s not true that either you should do X or you should do Y.” ↵
10. There is also non-normative research that falls under the label “decision theory,” which focuses on exploring the ways in which people do in practice make decisions or neutrally exploring the implications of different assumptions about decision-making processes. ↵
11. Arguably, even in academic literature, decision theories are often discussed under the implicit assumption that some form of subjectivism is true. However, it is also very easy to modify the theories to be compatible with theories that tell you to take into account things beyond your current desires. Value might be equated with one’s future welfare, for example, or with the total future welfare of all conscious beings. ↵
12. One thing that makes this issue a bit complicated is that rationalist community writing on decision theory sometimes seems to switch back and forth between describing decision theories as *normative claims about decisions* (which I believe is how academic philosophers typically describe decision theories) and as *algorithms to be used* (which seems to be inconsistent with how academic philosophers typically describe decision theories). I think this tendency to switch back and forth between describing decision theories in these two distinct ways can be seen both in papers proposing new decision theories and in online discussions. I also think this switching tendency can make things pretty confusing. Although it makes sense to discuss how an algorithm “performs”

when "implemented," once we specify a sufficiently precise performance metric, it does not seem to me to make sense to discuss the performance of a normative claim. I think the tendency to blur the distinction between algorithms and normative claims -- or, as Will MacAskill puts it in his recent and similar [critique](#), between "decision procedures" and "criteria of rightness" -- partly explains why proponents of FDT and other new decision theories have not been able to get much traction with academic decision theorists. For example, causal decision theorists are well aware that people who always take the actions that CDT says they should take will tend to fare less well in Newcomb scenarios than people who always take the actions that EDT says they should take. Causal decision theorists are also well aware that there are some scenarios -- for example, a Newcomb scenario with a perfect predictor and the option to get brain surgery to pre-commit yourself to one-boxing -- in which there is no available sequence of actions such that CDT says you should take each of the actions in the sequence. If you ask a causal decision theorist what sort of algorithm you should (according to CDT) put into an AI system that will live in a world full of Newcomb scenarios, if the AI system won't have the opportunity to self-modify, then I think it's safe to say a causal decision theorist won't tell you to put in an algorithm that only produces actions that CDT says it should take. This tells me that we really can't fluidly switch back and forth between making claims about the correctness of normative principles and claims about the performance of algorithms, as though there were a canonical one-to-one mapping between these two sorts of claims. Insofar as rationalist writing on decision theory tends to do this sort of switching, I suspect that it contributes to confusion on the part of many academic readers. See also this [blog post](#) by an academic decision theorist, Wolfgang Schwarz, for a much more thorough perspective on why proponents of FDT may be having difficulty getting traction within the academic decision theory community. ↩

13. A similar concern also leads me to assign low ($p < 10\%$) probability to normative decision theory research ultimately being useful for avoiding large-scale accidental harm caused by AI systems. It seems to me like the question "What is the correct decision theory?" only has an answer if we assume that realism is true. But even if we assume that realism is true, we are now asking a normative question ("What criterion determines whether an action is one an agent 'should' take?") as a way of trying to make progress on a non-normative question ("What approaches to designing advanced AI systems result in unintended disasters and which do not?"). Proponents of CDT and proponents of EDT do not actually disagree on how any given agent will behave, on what the causal outcome of assigning an agent a given algorithm will be, or on what evidence might be provided by the choice to assign an agent a given algorithm; they both agree, for example, about how much money different agents will tend to earn in the classic Newcomb scenario. What decision theorists appear to disagree about is a separate normative question that floats above (or rather "[supervenes](#)" upon) questions about observed behavior or questions about outcomes. I don't see how answering this normative question could help us much in answering the non-normative question of what approaches to designing advanced AI systems don't (e.g.) result in global catastrophe. Put another way, my concern is that the strategy here seems to rely on the hope that we can derive an "is" from an "ought."

However, in keeping with the above endnote, community work on decision theory only sometimes seems to be pitched (as it is in the abstract of [this paper](#)) as an exploration of normative principles. It is also sometimes pitched as an

exploration of how different “algorithms” “perform” across relevant scenarios. This exploration doesn't seem to me to have any direct link to the core academic decision theory literature and, given a sufficiently specific performance metric, does not seem to be inherently normative. I'm actually more optimistic, then, about this line of research having implications for AI development. Nonetheless, for reasons similar to the ones described in the post [“Decision Theory Anti-Realism,”](#) I'm still not very optimistic. In the cases that are being considered, the answer to the question “Which algorithm performs best?” will depend on subtle variations in the set of counterfactuals we consider when judging performance; different algorithms come out on top for different sets of counterfactuals. For example, in a prisoner's dilemma, the best-performing algorithm will vary depending on whether we are imaging a counterfactual world where just one agent was born with a different algorithm or a counterfactual world where both agents were born with different algorithms. It seems unclear to me where we go from here except perhaps to list several different sets of imaginary counterfactuals and note which algorithms perform best relative to them.

Wolfgang Schwarz and Will MacAskill also make similar points, regarding the sensitivity of comparisons of algorithmic performance, in their essays on FDT. Schwarz writes:

Yudkowsky and Soares constantly talk about how FDT "outperforms" CDT, how FDT agents "achieve more utility", how they "win", etc. As we saw above, it is not at all obvious that this is true. It depends, in part, on how performance is measured. At one place, Yudkowsky and Soares are more specific. Here they say that "in all dilemmas where the agent's beliefs are accurate [??] and the outcome depends only on the agent's actual and counterfactual behavior in the dilemma at hand -- reasonable constraints on what we should consider "fair" dilemmas -- FDT performs at least as well as CDT and EDT (and often better)". OK. But how we should we understand "depends on ... the dilemma at hand"? First, are we talking about subjunctive or evidential dependence? If we're talking about evidential dependence, EDT will often outperform FDT. And EDTers will say that's the right standard. CDTers will agree with FDTers that subjunctive dependence is relevant, but they'll insist that the standard Newcomb Problem isn't "fair" because here the outcome (of both one-boxing and two-boxing) depends not only on the agent's behavior in the present dilemma, but also on what's in the opaque box, which is entirely outside her control. Similarly for all the other cases where FDT supposedly outperforms CDT. Now, I can vaguely see a reading of "depends on ... the dilemma at hand" on which FDT agents really do achieve higher long-run utility than CDT/EDT agents in many "fair" problems (although not in all). But this is a very special and peculiar reading, tailored to FDT. We don't have any independent, non-question-begging criterion by which FDT always "outperforms" EDT and CDT across "fair" decision problems.

MacAskill writes:

[A]rguing that FDT does best in a class of ‘fair’ problems, without being able to define what that class is or why it's interesting, is a pretty weak argument. And, even if we could define such a class of cases, claiming that FDT ‘appears to be superior’ to EDT and CDT in the classic cases in the literature is simply begging the question: CDT adherents claims that two-boxing is the right action (which gets you more expected utility!) in

Newcomb's problem; EDT adherents claims that smoking is the right action (which gets you more expected utility!) in the smoking lesion. The question is which of these accounts is the right way to understand 'expected utility'; they'll therefore all differ on which of them do better in terms of getting expected utility in these classic cases.

↩

14. In my view, the epistemological issues are the most severe ones. I think Sharon Street's paper [A Darwinian Dilemma for Realist Theories of Value](#), for example, presents an especially hard-to-counter attack on the realist position on epistemological grounds. She argues that, in the light of the view that our brains evolved via natural selection, and natural selection did not and could not have directly selected for the accuracy of our normative intuitions, it is extremely difficult to construct a compelling explanation for why our normative intuitions should be correlated in any way with normative facts. This technically leave open the possibility of there being non-trivial normative facts, without us having any way of perceiving or intuiting them, but this state of affairs would strike most people as absurd. Although some realists, including Parfit, have attempted to counter Street's argument, I'm not aware of anyone who I feel has truly succeeded. Street's argument pretty much just seems to work to me. ↩
15. These metaphysical and epistemological issues become less concerning if we accept some version of "naturalist realism" which asserts that all normative claims can be reduced into claims about the natural world (i.e. claims about physical and psychological properties) and therefore tested in roughly the same way we might test any other claim about the natural world. However, this view seems wrong to me.

The bluntest objection to naturalist realism is what's sometimes called the "just-too-different" objection. This is the objection that, to many and perhaps most people, normative claims are just *obviously* a different sort of claim. No one has ever felt any inclination to evoke an "is/is-made-of-wood divide" or an "is/is-illegal-in-Massachusetts divide," because the property of being made of wood and the property of being illegal in Massachusetts are obviously properties of the standard (natural) kind. But references to the "is/ought divide" -- or, equivalently, the distinction between the "positive" and the "normative" -- are commonplace and don't typically provoke blank stares. Normative discussions are, seemingly, about something *above-and-beyond* and *distinct from* discussions of the physical and psychological aspects of a situation. When people debate whether or not it's "wrong" to support the death penalty or "wrong" for women to abort unwanted pregnancies, for example, it seems obvious that physical and psychological facts are typically not the core (or at least *only*) thing in dispute.

G.E. Moore's "[Open Question Argument](#)" elaborates on this objection. The argument also raises the point that that, in many cases where we are inclined to ask "What should I do?", it seems like what we are inclined to ask goes above-and-beyond any individual question we might ask about the natural world. Consider again the case where we are considering a career change and wondering what we should do. It seems like we could know all of the natural facts -- facts like how happy will we be on average while pursuing each career, how satisfied will we feel looking back on each career, how many lives we could improve by donating money made in each career, what labor practices each

company has, how disappointed our parents will be if we pursue each career, how our personal values will change if we pursue each career, what we would end up deciding at the end of one hypothetical deliberative process or another, etc. -- and still retain the inclination to ask, "Given all this, what should I do?" This means that -- insofar as we're taking the realist stance that this question actually has a meaningful answer, rather than rejecting the question as [vacuous](#) -- the claim that we "should" do one thing or another cannot easily be understood as a claim about the natural world. A set of claims about the natural world may *support* the claim that we should make a certain decision, but, in cases such as this one, it seems like no set of claims about the natural world is *equivalent* to the claim that we should make a certain decision.

A last objection to mention is Parfit's "Triviality Objection" (*On What Matters*, Section 95). The basic intuition behind Parfit's objection is that pretty much any attempt to define the word "should" in terms of natural properties would turn many normative claims into puzzling assertions of either obvious tautologies or obvious falsehoods. For example, consider a man who is offered -- at the end of his life, I guess by the devil or something -- the option of undergoing a year of certain torture for a one-in-a-trillion chance of receiving a big prize: a trillion years of an equivalently powerful positive experience, plus a single lollipop. He is purely interested in experiencing pleasure and avoiding pain and would like to know whether he should take the offer. A decision theorist who endorses expected desire-fulfillment maximisation says that he "should," since the lollipop tips the offer over into having slightly positive expected value. A decision theorist who endorses risk aversion says he "should not," since the man is nearly certain to be horribly tortured without receiving any sort of compensation. In this context, it's hard to understand how we could redefine the claim "He should take action X" in terms of natural properties and have this disagreement make any sense. We could define the phrase as meaning "Action X maximizes expected fulfillment of desire," but now the first decision theorist is expressing an obvious tautology and the second decision theorist is expressing an obvious falsehood. We could also try, in keeping with a [suggestion](#) by Eliezer Yudkowsky, to define the phrase as meaning "Action X is the one that someone acting in a winning way would take." But this is obviously too vague to imply a particular action; taking the gamble is associated with some chance of winning and some chance of losing. We could make the definition more specific -- for instance, saying "Action X is the one that someone acting in a way that maximizes expected winning would take" -- but now of course we're back in tautology mode. The apparent upshot, here, is that many normative claims simply can't be interpreted as non-trivially true or non-trivially false claims about natural properties. The associated disagreements only become sensible if we interpret them as being about something above-and-beyond these properties.

Of course, it is surely true that *some* of the claims people make using the word "should" can be understood as claims about the natural world. Words can, after all, be used in many different ways. But it's the claims that can't easily be understood in this way that non-naturalist realists such as Parfit, Enoch, and Moore have in mind. In general, I agree with the view that the key division in metaethics is between self-identified non-naturalist realists on the one hand and self-identified anti-realists and naturalist realists on the other hand, since "naturalist realists" are in fact anti-realists with regard to the distinctively normative properties of decisions that non-naturalist realists are talking about. If we rule out non-naturalist realism as a position then it seems the main remaining question is a somewhat boring one about semantics: When someone

makes a statement of form “A should do X,” are they most commonly expressing some sort of attitude (non-cognitivism), making a claim about the natural world (naturalist realism), or making a claim about some made-up property that no actions actually possess (error theory)?

Here, for example, is how Michael Huemer (a non-naturalist realist) expresses this point in his book *Ethical Intuitionism* (pg. 8):

[Non-naturalist realists] differ fundamentally from everyone else in their view of the world. [Naturalist realists], non-cognitivists, and nihilists all agree in their basic view of the world, for they have no significant disagreements about what the non-evaluative facts are, and they all agree that there are no further facts over and above those. They agree, for example, on the non-evaluative properties of the act of stealing, and they agree, contra the [non-naturalist realists], that there is no further, distinctively evaluative property of the act. Then what sort of dispute do the [three] monistic theories have? I believe that, though this is not generally recognized, their disputes with each other are merely semantic. Once the nature of the world ‘out there’ has been agreed upon, semantic disputes are all that is left.

I think this attitude is in line with the viewpoint that Luke Muehlhauser expresses in his classic LessWrong blog post on what he calls “[pluralistic moral reductionism](#).” PMR seems to me to be the view that: (a) non-naturalist realism is false, (b) all remaining meta-normative disputes are purely semantic, and (c) purely semantic disputes aren’t terribly substantive and often reflect a failure to accept that the same phrase can be used in different ways. If we define the view this way, then, *conditional on non-naturalist realism being false*, I believe that PMR is the correct view. I believe that many non-naturalist realists would agree on this point as well. ↩

16. This point is made by Parfit in *On What Matters*. He writes: “We could not have decisive reasons to believe that there are no such normative truths, since the fact that we had these reasons would itself have to be one such truth. This point may not refute this kind of skepticism, since some skeptical arguments might succeed even if they undermined themselves. But this point shows how deep such skepticism goes, and how blank this skeptical state of mind would be” (*On What Matters*, Section 86). ↩
17. The [PhilPapers survey](#) suggests that philosophers who favor realism outweigh philosophers who favor anti-realism by about a 2:1 ratio. ↩

Seven habits towards highly effective minds

Lately I've been thinking about how my thinking works, and how it can be improved. The simplest way to do so is probably to nudge myself towards paying more attention to various useful habits of mind. Here are the ones I've found most valuable (roughly in order):

1. Tying together the act of saying a statement, and the act of evaluating whether I actually believe it. After making a novel claim, saying out loud to myself: "is this actually true?" and "how could I test this?"
2. Being comfortable with pausing to reflect and thinking out loud. Trying to notice when my responses are too quick and reflexive, as a sign that I'm not thinking hard enough about the point I'm addressing.
3. Asking for [specific examples](#), and using more of my own. Tabooing vague abstractions and moving away from discussing claims that are too general.
4. Being charitable and collaborative, both towards new ideas and towards conversational partners. Trying to rephrase other people's arguments and pass [Ideological Turing Tests](#) on them. Helping my conversational partners build up their ideas.
5. Noticing the affect heuristic, and which claims stir up emotions. Noticing when I'm talking defensively or heatedly, and when it'd be uncomfortable to believe something.
6. Thinking in terms of probabilities; cashing out beliefs in terms of predictions; then betting on them. I haven't done enough bets to calibrate myself well, but I find that even just the feeling of having money on the line is often enough to make me rethink. Being asked whether something is a crux gives me a similar feeling.
7. Thinking about how the conversations and debates I participate in actually create value, and when they should be redirected or halted.

Then there are social influences. I think one of the greatest virtues of the rationalist community is in creating an environment which encourages the use of the tools above. Another example: my girlfriend fairly regularly points out times when I've contradicted myself. I think this has helped me notice and limit the extent to which I behave like an opinion confabulation machine.

I'd classify most if not all of the tools listed above as [tools for evaluating ideas](#), though, rather than tools for generating ideas. What helps with the latter? I've personally found that one very useful strategy is to make and then justify bold claims based on vague intuitions. In the process of defending my position, I'm forced to actually flesh it out and make it coherent (although I do need to be careful not to become overly attached to the untrue parts). And what's helped the most is that after having interesting conversations, I now write posts inspired by them much more frequently. I often feel like Feynman in [this story](#): "When historian Charles Weiner found pages of Nobel Prize-winning physicist Richard Feynman's notes, he saw it as a "record" of Feynman's work. Feynman himself, however, insisted that the notes were not a record but the work itself." Arguing and writing are not just ways to transmit my thoughts, but also the key mechanisms by which I generate new thoughts.

(Edited to add: a friend pointed out that the last line is a good indicator that I'm being insufficiently empirical. I think I agree; it should also include the mechanism of looking at the world and noticing something confusing going on.)

Focus

One of the things I have the hardest time with is focus. I have times when focus comes easily, and I get a week's worth of work done in hours, come up with ideas that need sustained thinking, write blog posts that have been bouncing around my head for years. Other times my attention is flighty, and I have to struggle to keep myself from finding low-investment sources of entertainment.

The biggest component of this is how excited I am. When I really get into something, that's when focus comes easily. I recently finished a project at work where I was building something that dramatically improved something the team had found frustrating for years. I could see where I wanted it to go, and I was excited enough that when I ran into issues I pushed at it until I had good solutions for them. The [bass whistle](#) project was another one like this, where I couldn't think of anything else for about a week, until I had something coded up and working. I didn't need to make myself focus, I needed to make myself do the rest of my life.

A lot of other factors matter, but are mediated by excitement. Short iteration cycles, where I can quickly find out whether something worked, are so important to me because they keep the excitement from draining away. Doing something no one has done before, that a lot of people are going to like, that needs doing urgently, or that I've been thinking about for a long time all help, but mostly because those are exciting things.

In-person collaboration also helps, in at least two ways. When I'm working with someone, talking to them directly, it feels like ideas flow much better. It's most fun when we have complementary skills, each filling in for and learning from the other, but even when the other person is inexperienced it still helps to have another person's worth of working memory. And then if I'm working one-on-one with someone I can't respond to brief roadblocks by letting myself get distracted.

There's also a component of mental patterns: if I'm in the habit of tabbing over to Facebook I'll fall out of focus more easily. [1] The hard part for me is that I'm often doing work that is full of short breaks that should be fine to fill: waiting for compiles, for queries, for tests. This means that when I've tried to make myself rules like "no distraction activities" I either get bored enough waiting for things that I can't stick to the rule, or I learn how to turn some previously fine activity into a diversion. The feeling of "I can't make progress right now, let's distract" is shared between "my code's compiling", where waiting will help, and "this problem is hard", where waiting (mostly) won't. How well I'm able to distinguish these in the moment varies, however, and I'm not all that good at it. This also means that if I'm doing a kind of work that lends itself to unbroken effort (washing dishes, framing a wall, coding something that's fully in my head) then I'm much more likely to just work until I'm done.

There are also kinds of work where I need to be in a very distraction-prone mindset. Analysis is often like this for me, where I relax my barrier between having an idea and trying it. Sometimes this leads me to explore aspects of a problem that are really promising, and other times it leads me to explore the history of bi-level railcars. You would think this would be very easy to reign in, and it is for me when I'm excited about the analysis, but my excitement can go off in pretty random directions.

Another aspect is that when I have something I have to do that I'm just not excited about, it's really hard to get myself to do it. I do have strategies, like sitting down with just a piece of paper or a single browser tab and telling myself I can't get up until I've hit some criteria, but it goes very slowly and is very unpleasant. Sometimes leaving tasks like this until I do get excited about them helps, but not if I never end up feeling that way.

In the other direction, I've done some of my best work while distracted from something else that I was supposed to be doing: my sense of what's exciting sometimes gets at something my conscious prioritization doesn't. Additionally, "doing things when I'm excited about them" often means "do things when they're most tractable". I'm nervous about breaking something that overall has outcomes I like.

Some of this feels like what Constantin describes in [The Costs of Reliability](#). The whole post is good, but in particular: "people given an open-ended mandate to do what they like can be far more efficient than people working to spec... at the cost of unpredictable output with no guarantees of getting what you need when you need it".

There's also a lot that resonates in Graham's [Disconnecting Distraction](#), including "Another reason it was hard to notice the danger of this new type of distraction was that social customs hadn't yet caught up with it. If I'd spent a whole morning sitting on a sofa watching TV, I'd have noticed very quickly. That's a known danger sign, like drinking alone. But using the Internet still looked and felt a lot like work." Time holes and culture of handling them are both evolving, and figuring out how to keep from falling into them involves constantly learning how to tell good from bad uses.

Overall, this both is and isn't a big problem for me. It isn't, in that I am often very focused and get a lot done, at work and at home. It is, in a sense of opportunity cost: possibly I could be doing a lot more if I were able to focus more by either (a) influencing my excitement or (b) doing important things without excitement.

I initially wrote this as a "here's where I am" post, but rereading it I do have a few ideas:

- Putting more effort into recognizing the cases where "wait a bit" will help me make progress
- Bring a book to work to read during times when waiting really is the best next step, since it's both much less addictive than a digital device and better understood.
- Spending more time working directly with others.

[1] A strong correlate for this with me is, if I put a piece of chocolate in my mouth, do I chew it or am I able to wait and enjoy it slowly? And, weirdly, this seems to still give signal even though I now use it for self-evaluation.

Comment via: [facebook](#)

A Critique of Functional Decision Theory

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

A Critique of Functional Decision Theory

NB: My writing this note was prompted by Carl Shulman, who suggested we could try a low-time-commitment way of attempting to understand the disagreement between some folks in the rationality community and academic decision theorists (including myself, though I'm not much of a decision theorist). Apologies that it's sloppier than I'd usually aim for in a philosophy paper, and lacking in appropriate references. And, even though the paper is pretty negative about FDT, I want to emphasise that my writing this should be taken as a sign of respect for those involved in developing FDT. I'll also caveat I'm unlikely to have time to engage in the comments; I thought it was better to get this out there all the same rather than delay publication further.

1. Introduction

There's a long-running issue where many in the rationality community take functional decision theory (and its variants) very seriously, but the academic decision theory community does not. But there's been little public discussion of FDT from academic decision theorists (one exception is [here](#)); this note attempts to partly address this gap.

So that there's a clear object of discussion, I'm going to focus on Yudkowsky and Soares' '[Functional Decision Theory](#)' (which I'll refer to as Y&S), though I also read a revised version of Soares and Levinstein's [Cheating Death in Damascus](#).

This note is structured as follows. Section II describes causal decision theory (CDT), evidential decision theory (EDT) and functional decision theory (FDT). Sections III-VI describe problems for FDT: (i) that it sometimes makes bizarre recommendations, recommending an option that is certainly lower-utility than another option; (ii) that it fails to one-box in most instances of Newcomb's problem, even though the correctness of one-boxing is supposed to be one of the guiding motivations for the theory; (iii) that it results in implausible discontinuities, where what is rational to do can depend on arbitrarily small changes to the world; and (iv) that, because there's no real fact of the matter about whether a particular physical process implements a particular algorithm, it's deeply indeterminate what FDT's implications are. In section VII I discuss the idea that FDT 'does better at getting utility' than EDT or CDT; I argue that Y&S's claims to this effect are unhelpfully vague, and on any more precise way of understanding their claim, aren't plausible. In section VIII I briefly describe a view that captures some of the motivation behind FDT, and in my view is more plausible. I conclude that FDT faces a number of deep problems and little to say in its favour.

In what follows, I'm going to assume a reasonable amount of familiarity with the debate around Newcomb's problem.

II. CDT, EDT and FDT

Informally; CDT, EDT and FDT differ in what non-causal correlations they care about when evaluating a decision. For CDT, what you cause to happen is all that matters; if your action correlates with some good outcome, that's nice to know, but it's not relevant to what you ought to do. For EDT, all correlations matter: you should pick whatever action will result in you believing you will have the highest expected utility. For FDT, only some non-causal correlations matter, namely only those correlations between your action and events elsewhere in time and space that would be different in the (logically impossible) worlds in which the output of the algorithm you're running is different. Other than for those correlations, FDT behaves in the same way as CDT.

Formally, where S represents states of nature, A, B etc represent acts, P is a probability function, and $U(S_i \& A)$ represents the utility the agent gains from the outcome of choosing A given state S_i , and ' \geq ' represents the 'at least as choiceworthy as' relation:

On EDT:

$$A \geq B \text{ iff } \sum P(S_i | A) U(S_i \& A) \geq \sum P(S_i | B) U(S_i \& B)$$

Where ' $|$ ' represents conditional probability.

On CDT:

$$A \geq B \text{ iff } \sum P(S_i \setminus A) U(S_i \& A) \geq \sum P(S_i \setminus B) U(S_i \& B)$$

Where ' \setminus ' is a 'causal probability function' that represents the decision-maker's judgments about her ability to causally influence the events in the world by doing a particular action. Most often, this is interpreted in counterfactual terms (so $P(S \setminus A)$ represents something like the probability of S coming about were I to choose A) but it needn't be.

On FDT:

$$A \geq B \text{ iff } \sum P(S_i \dagger A) U(S_i \& A) \geq \sum P(S_i \dagger B) U(S_i \& B)$$

Where I introduce the operator " \dagger " to represent the special sort of function that Yudkowsky and Soares propose, where $P(S \dagger A)$ represents the probability of S occurring were the output of the algorithm that the decision-maker is running, in this decision situation, to be A . (I'm not claiming that it's clear what this means. E.g. see here, second bullet point, arguing there can be no such probability function, because any probability function requires certainty in logical facts and all their entailments. I also note that strictly speaking FDT doesn't assess acts in the same sense that CDT assesses acts; rather it assesses algorithmic outputs, and that Y&S have a slightly different formal set up than the one I describe above. I don't think this will matter for the purposes of this note, though.)

With these definitions on board, we can turn to objections to FDT.

III. FDT sometimes makes bizarre recommendations

The criterion that Y&S regard as most important in assessing a decision theory is ‘amount of utility achieved’. I think that this idea is importantly underspecified (which I discuss more in section VII), but I agree with the spirit of it. But FDT does very poorly by that criterion, on any precisification of it.

In particular, consider the following principle:

Guaranteed Payoffs: In conditions of certainty — that is, when the decision-maker has no uncertainty about what state of nature she is in, and no uncertainty about the utility payoff of each action is — the decision-maker should choose the action that maximises utility.

That is: for situations where there’s no uncertainty, we don’t need to appeal to expected utility theory in any form to work out what we ought to do. You just ought to do whatever will give you the highest utility payoff. This should be a constraint on any plausible decision theory. But FDT violates that principle.

Consider the following case:

Bomb.

You face two open boxes, Left and Right, and you must take one of them. In the Left box, there is a live bomb; taking this box will set off the bomb, setting you ablaze, and you certainly will burn slowly to death. The Right box is empty, but you have to pay \$100 in order to be able to take it.

A long-dead predictor predicted whether you would choose Left or Right, by running a simulation of you and seeing what that simulation did. If the predictor predicted that you would choose Right, then she put a bomb in Left. If the predictor predicted that you would choose Left, then she did not put a bomb in Left, and the box is empty.

The predictor has a failure rate of only 1 in a trillion trillion. Helpfully, she left a note, explaining that she predicted that you would take Right, and therefore she put the bomb in Left.

You are the only person left in the universe. You have a happy life, but you know that you will never meet another agent again, nor face another situation where any of your actions will have been predicted by another agent. What box should you choose?

The right action, according to FDT, is to take Left, in the full knowledge that as a result you will slowly burn to death. Why? Because, using Y&S’s counterfactuals, *if* your algorithm were to output ‘Left’, then it would also have outputted ‘Left’ when the predictor made the simulation of you, and there would be no bomb in the box, and you could save yourself \$100 by taking Left. In contrast, the right action on CDT or EDT is to take Right.

The recommendation is implausible enough. But if we stipulate that in this decision-situation the decision-maker is certain in the outcome that her actions would bring about, we see that FDT violates *Guaranteed Payoffs*.

(One might protest that no good Bayesian would ever have credence 1 in an empirical proposition. But, first, that depends on what we could call ‘evidence’ — if a proposition is part of your evidence base, you have credence 1 in it. And, second, we could construct very similar principles to *Guaranteed Payoffs* that don’t rely on the idea of certainty, but on approximations to certainty.)

Note that FDT’s recommendation in this case is *much* more implausible than even the worst of the *prima facie* implausible recommendations of EDT or CDT. So, if we’re going by appeal to cases, or by ‘who gets more utility’, FDT is looking very unmotivated.

IV. FDT fails to get the answer Y&S want in most instances of the core example that’s supposed to motivate it

On FDT, you consider what things would look like in the closest (logically impossible) world in which the algorithm you are running were to produce a different output than what it in fact does. Because, so the argument goes, in Newcomb problems the predictor is also running your algorithm, or a ‘sufficiently similar’ algorithm, or a representation of your algorithm, you consider the correlation between your action and the predictor’s prediction (even though you don’t consider other sorts of correlations.)

However, the predictor needn’t be running your algorithm, or have anything like a representation of that algorithm, in order to predict whether you’ll one box or two-box. Perhaps the Scots tend to one-box, whereas the English tend to two-box. Perhaps the predictor knows how you’ve acted prior to that decision. Perhaps the Predictor painted the transparent box green, and knows that’s your favourite colour and you’ll struggle not to pick it up. In none of these instances is the Predictor plausibly doing anything like running the algorithm that you’re running when you make your decision. But they are still able to predict what you’ll do. (And bear in mind that the Predictor doesn’t even need to be very reliable. As long as the Predictor is better than chance, a Newcomb problem can be created.)

In fact, on the vast majority of ways that the Predictor could predict your behavior, she isn’t running the algorithm that you are running, or representing it. But if the Predictor isn’t running the algorithm that you are running, or representing it, then, on the most natural interpretation, FDT will treat this as ‘mere statistical correlation’, and therefore act like CDT. So, in the vast majority of Newcomb cases, FDT would recommend two-boxing. But the intuition in favour of one-boxing in Newcomb cases was exactly what was supposed to motivate FDT in the first place.

Could we instead interpret FDT, such that it doesn’t have to require the Predictor to be running the exact algorithm — some similar algorithm would do? But I’m not sure how that would help: in the examples given above, the Predictor’s predictions aren’t based on anything like running your algorithm. In fact, the predictor may know very little about you, perhaps only whether you’re English or Scottish.

One could suggest that, even though the Predictor is not running a sufficiently similar algorithm to you, nonetheless the Predictor’s prediction is subjunctively dependent on

your decision (in the Y&S sense of ‘subjunctive’). But, without any account of Y&S’s notion of subjunctive counterfactuals, we just have no way of assessing whether that’s true or not. Y&S note that specifying an account of their notion of counterfactuals is an ‘open problem,’ but the problem is much deeper than that. Without such an account, it becomes completely indeterminate what follows from FDT, even in the core examples that are supposed to motivate it — and that makes FDT not a new decision theory so much as a promissory note.

Indeed, on the most plausible ways of cashing this out, it doesn’t give the conclusions that Y&S would want. If I imagine the closest world in which $6288 + 1048 = 7336$ is false (Y&S’s example), I imagine a world with laws of nature radically unlike ours — because the laws of nature rely, fundamentally, on the truths of mathematics, and if one mathematical truth is false then either (i) mathematics as a whole must be radically different, or (ii) all mathematical propositions are true because it is simple to prove a contradiction and every proposition follows from a contradiction. Either way, when I imagine worlds in which FDT outputs something different than it in fact does, then I imagine valueless worlds (no atoms or electrons, etc) — and this isn’t what Y&S are wanting us to imagine.

Alternatively (as Abram Demski suggested to me in a comment), Y&S could accept that the decision-maker *should* two-box in the cases given above. But then, it seems to me, that FDT has lost much of its initial motivation: the case for one-boxing in Newcomb’s problem didn’t seem to stem from whether the Predictor was running a simulation of me, or just using some other way to predict what I’d do.

V. Implausible discontinuities

A related problem is as follows: FDT treats ‘mere statistical regularities’ very differently from predictions. But there’s no sharp line between the two. So it will result in implausible discontinuities. There are two ways we can see this.

First, take some physical processes S (like the lesion from the Smoking Lesion) that causes a ‘mere statistical regularity’ (it’s not a Predictor). And suppose that the existence of S tends to cause both (i) one-boxing tendencies and (ii) whether there’s money in the opaque box or not when decision-makers face Newcomb problems. If it’s S alone that results in the Newcomb set-up, then FDT will recommending two-boxing.

But now suppose that the pathway by which S causes there to be money in the opaque box or not is that another agent looks at S and, if the agent sees that S will cause decision-maker X to be a one-boxer, then the agent puts money in X’s opaque box. Now, because there’s an agent making predictions, the FDT adherent will presumably want to say that the right action is one-boxing. But this seems arbitrary — why should the fact that S’s causal influence on whether there’s money in the opaque box or not go via another agent much such a big difference? And we can think of all sorts of spectrum cases in between the ‘mere statistical regularity’ and the full-blooded Predictor: What if the ‘predictor’ is a very unsophisticated agent that doesn’t even understand the implications of what they’re doing? What if they only partially understand the implications of what they’re doing? For FDT, there will be some point of sophistication at which the agent moves from simply being a conduit for a causal process to instantiating the right sort of algorithm, and suddenly FDT will switch from recommending two-boxing to recommending one-boxing.

Second, consider that same physical process S, and consider a sequence of Newcomb cases, each of which gradually make S more and more complicated and agent-y, making it progressively more similar to a Predictor making predictions. At some point, on FDT, there will be a point at which there's a sharp jump; prior to that point in the sequence, FDT would recommend that the decision-maker two-boxes; after that point, FDT would recommend that the decision-maker one-boxes. But it's very implausible that there's some S such that a tiny change in its physical makeup should affect whether one ought to one-box or two-box.

VI. FDT is deeply indeterminate

Even putting the previous issues aside, there's a fundamental way in which FDT is indeterminate, which is that there's no objective fact of the matter about whether two physical processes A and B are running the same algorithm or not, and therefore no objective fact of the matter of which correlations represent implementations of the same algorithm or are 'mere correlations' of the form that FDT wants to ignore. (Though I'll focus on 'same algorithm' cases, I believe that the same problem would affect accounts of when two physical processes are running similar algorithms, or any way of explaining when the output of some physical process, which instantiates a particular algorithm, is Y&S-subjunctively dependent on the output of another physical process, which instantiates a different algorithm.)

To see this, consider two calculators. The first calculator is like calculators we are used to. The second calculator is from a foreign land: it's identical except that the numbers it outputs always come with a negative sign ('-') in front of them when you'd expect there to be none, and no negative sign when you expect there to be one. Are these calculators running the same algorithm or not? Well, perhaps on this foreign calculator the '-' symbol means what we usually take it to mean — namely, that the ensuing number is negative — and therefore every time we hit the '=' button on the second calculator we are asking it to run the algorithm 'compute the sum entered, then output the negative of the answer'. If so, then the calculators are systematically running different algorithms.

But perhaps, in this foreign land, the '-' symbol, in this context, means that the ensuing number is positive and the lack of a '-' symbol means that the number is negative. If so, then the calculators are running exactly the same algorithms; their differences are merely notational.

Ultimately, in my view, all we have, in these two calculators, are just two physical processes. The further question of whether they are running the same algorithm or not depends on how we interpret the physical outputs of the calculator. There is no deeper fact about whether they're 'really' running the same algorithm or not. And in general, it seems to me, there's no fact of the matter about which algorithm a physical process is implementing in the absence of a particular interpretation of the inputs and outputs of that physical process.

But if that's true, then, even in the Newcomb cases where a Predictor is simulating you, it's a matter of choice of symbol-interpretation whether the predictor ran the same algorithm that you are now running (or a representation of that same algorithm). And the way you choose that symbol-interpretation is fundamentally arbitrary. So there's no real fact of the matter about whether the predictor is running the same algorithm as you. It's indeterminate how you should act, given FDT: you

should one-box, given one way of interpreting the inputs and outputs of the physical process the Predictor is running, but two-box given an alternative interpretation.

Now, there's a [bunch of interesting work](#) on concrete computation, trying to give an account of when two physical processes are performing the same computation. The best response that Y&S could make to this problem is to provide a compelling account of when two physical processes are running the same algorithm that gives them the answers they want. But almost all accounts of computation in physical processes have the issue that very many physical processes are running very many different algorithms, all at the same time. (Because most accounts rely on there being some mapping from physical states to computational states, and there can be multiple mappings.) So you might well end up with the problem that in the closest (logically impossible) world in which FDT outputs something other than what it does output, not only do the actions of the Predictor change, but so do many other aspects of the world. For example, if the physical process underlying some aspect of the US economy just happened to be isomorphic with FDT's algorithm, then in the logically impossible world where FDT outputs a different algorithm, not only does the predictor act differently, but so does the US economy. And that will probably change the value of the world under consideration, in a way that's clearly irrelevant to the choice at hand.

VII. But FDT gets the most utility!

Y&S regard the most important criterion to be 'utility achieved', and thinks that FDT does better than all its rivals in this regard. Though I agree with something like the spirit of this criterion, its use by Y&S is unhelpfully ambiguous. To help explain this, I'll go on a little detour to present some distinctions that are commonly used by academic moral philosophers and, to a lesser extent, decision theorists. (For more on these distinctions, see Toby Ord's DPhil thesis.)

Evaluative focal points

An *evaluative focal point* is an object of axiological or normative evaluation. ('Axiological' means 'about goodness/badness'; 'normative' means 'about rightness/wrongness'. If you're a consequentialist, x is best iff it's right, but if you're a non-consequentialist the two can come apart.) When doing moral philosophy or decision theory, the most common evaluative focal points are *acts*, but we can evaluate other things too: characters, motives, dispositions, sets of rules, beliefs, and so on.

Any axiological or normative theory needs to specify which focal point it is evaluating. The theory can evaluate a single focal point (e.g. act utilitarianism, which only evaluates acts) or many (e.g. global utilitarianism, which evaluates everything).

The theory can also differ on whether it is *direct* or *indirect* with respect to a given evaluative focal point. For example, Hooker's rule-consequentialism is a direct theory with respect to sets of rules, and an indirect theory with respect to acts: it evaluates sets of rules on the basis of their consequences, but evaluates acts with respect to how they conform to those sets of rules. Because of this, on Hooker's view, the right act need not maximize good consequences.

Criterion of rightness vs decision procedure

In chess, there's a standard by which it is judged who has won the game, namely, the winner is whoever first puts their opponent's king into checkmate. But relying solely on that standard of evaluation isn't going to go very well if you actually want to win at chess. Instead, you should act according to some other set of rules and heuristics, such as: "if white, play e4 on the first move," "don't get your Queen out too early," "rooks are worth more than bishops" etc.

A similar distinction can be made for axiological or normative theories. The criterion of rightness, for act utilitarianism, is, "The right actions are those actions which maximize the sum total of wellbeing." But that's not the decision procedure one ought to follow. Instead, perhaps, you should rely on rules like 'almost never lie', 'be kind to your friends and family', 'figure out how much you can sustainably donate to effective charities, and do that,' and so on.

For some people, in fact, learning that utilitarianism is true will cause one to be a worse utilitarian by the utilitarian's criterion of rightness! (Perhaps you start to come across as someone who uses others as means to an end, and that hinders your ability to do good.) By the utilitarian criterion of rightness, someone could in principle act rightly in every decision, even though they have never heard of utilitarianism, and therefore never explicitly tried to follow utilitarianism.

These distinctions and FDT

From Y&S, it wasn't clear to me whether FDT is really meant to assess acts, agents, characters, decision procedures, or outputs of decision procedures, and it wasn't clear to me whether it is meant to be a direct or an indirect theory with respect to acts, or with respect to outputs of decision procedures. This is crucial, because it's relevant to which decision theory 'does best at getting utility'.

With these distinctions in hand, we can see that Y&S employ multiple distinct interpretations of their key criterion. Sometimes, for example, Y&S talk about how "FDT agents" (which I interpret as 'agents who follow FDT to make decisions') get more utility, e.g.:

- "Using one simple and coherent decision rule, functional decision theorists (for example) achieve more utility than CDT on Newcomb's problem, more utility than EDT on the smoking lesion problem, and more utility than both in Parfit's hitchhiker problem."
- "We propose an entirely new decision theory, functional decision theory (FDT), that maximizes agents' utility more reliably than CDT or EDT."
- "FDT agents attain high utility in a host of decision problems that have historically proven challenging to CDT and EDT: FDT outperforms CDT in Newcomb's problem; EDT in the smoking lesion problem; and both in Parfit's hitchhiker problem."
- "It should come as no surprise that an agent can outperform both CDT and EDT as measured by utility achieved; this has been known for some time (Gibbard and Harper 1978)."
- "Expanding on the final argument, proponents of EDT, CDT, and FDT can all agree that it would be great news to hear that a beloved daughter adheres to FDT, because FDT agents get more of what they want out of life. Would it not then be strange if the correct theory of rationality were some alternative to the theory that produces the best outcomes, as measured in utility? (Imagine hiding decision theory textbooks from loved ones, lest they be persuaded to adopt the "correct" theory and do worse thereby!) We consider this last argument—the

argument from utility—to be the one that gives the precommitment and value-of-information arguments their teeth. If self-binding or self-blinding were important for getting more utility in certain scenarios, then we would plausibly endorse those practices. Utility has primacy, and FDT's success on that front is the reason we believe that FDT is a more useful and general theory of rational choice.”

Sometimes Y&S talk about how different decision theories produce more utility on average if they were to face a specific dilemma repeatedly:

- “Measuring by utility achieved on average over time, CDT outperforms EDT in some well-known dilemmas (Gibbard and Harper 1978), and EDT outperforms CDT in others (Ahmed 2014b).”
- “Imagine an agent that is going to face first Newcomb’s problem, and then the smoking lesion problem. Imagine measuring them in terms of utility achieved, by which we mean measuring them by how much utility we expect them to attain, on average, if they face the dilemma repeatedly. The sort of agent that we’d expect to do best, measured in terms of utility achieved, is the sort who one-boxes in Newcomb’s problem, and smokes in the smoking lesion problem.”

Sometimes Y&S talk about which agent will achieve more utility ‘in expectation’, though they don’t define the point at which they gain more expected utility (or what notion of ‘expected utility’ is being used):

- “One-boxing in the transparent Newcomb problem may look strange, but it works. Any predictor smart enough to carry out the arguments above can see that CDT and EDT agents two-box, while FDT agents one-box. Followers of CDT and EDT will therefore almost always see an empty box, while followers of FDT will almost always see a full one. Thus, FDT agents achieve more utility in expectation.”

Sometimes they talk about how much utility ‘decision theories tend to achieve in practice’:

- “It is for this reason that we turn to Newcomblike problems to distinguish between the three theories, and demonstrate FDT’s superiority, when measuring in terms of utility achieved.”
- “we much prefer to evaluate decision theories based on how much utility they tend to achieve in practice.”

Sometimes they talk about how well the decision theory does in a circumscribed class of cases (though they note in footnote 15 that they can’t define what this class of cases are):

- “FDT does appear to be superior to CDT and EDT in all dilemmas where the agent’s beliefs are accurate and the outcome depends only on the agent’s behavior in the dilemma at hand. Informally, we call these sorts of problems “fair problems.””
- “FDT, we claim, gets the balance right. An agent who weighs her options by imagining worlds where her decision function has a different output, but where logical, mathematical, nomic, causal, etc. constraints are otherwise respected, is an agent with the optimal predisposition for whatever fair dilemma she encounters.”

And sometimes they talk about how much utility the agent would receive in different possible worlds than the one she finds herself in:

- “When weighing actions, Fiona simply imagines hypotheticals corresponding to those actions, and takes the action that corresponds to the hypothetical with higher expected utility—even if that means imagining worlds in which her observations were different, and even if that means achieving low utility in the world corresponding to her actual observations.”

As we can see, the most common formulation of this criterion is that they are looking for the decision theory that, if *run by an agent*, will produce the most utility over their lifetime. That is, they’re asking what the best *decision procedure* is, rather than what the best *criterion of rightness* is, and are providing an *indirect* account of the rightness of acts, assessing acts in terms of how well they conform with the best decision procedure.

But, if that’s what’s going on, there are a whole bunch of issues to dissect. First, it means that FDT is not playing the same game as CDT or EDT, which are proposed as criteria of rightness, directly assessing acts. So it’s odd to have a whole paper comparing them side-by-side as if they are rivals.

Second, what decision theory does best, if run by an agent, depends crucially on what the world is like. To see this, let’s go back to question that Y&S ask of what decision theory I’d want my child to have. This depends on a whole bunch of empirical facts: if she might have a gene that causes cancer, I’d hope that she adopts EDT; though if, for some reason, I knew whether or not she did have that gene and she didn’t, I’d hope that she adopts CDT. Similarly, if there were long-dead predictors who can no longer influence the way the world is today, then, if I didn’t know what was in the opaque boxes, I’d hope that she adopts EDT (or FDT); if I did know what was in the opaque boxes (and she didn’t) I’d hope that she adopts CDT. Or, if I’m in a world where FDT-ers are burned at the stake, I’d hope that she adopts anything other than FDT.

Third, the best decision theory to run is not going to look like *any* of the standard decision theories. I don’t run CDT, or EDT, or FDT, and I’m very glad of it; it would be impossible for my brain to handle the calculations of any of these decision theories every moment. Instead I almost always follow a whole bunch of rough-and-ready and much more computationally tractable heuristics; and even on the rare occasions where I do try to work out the expected value of something explicitly, I don’t consider the space of all possible actions and all states of nature that I have some credence in — doing so would take years.

So the main formulation of Y&S’s most important principle doesn’t support FDT. And I don’t think that the other formulations help much, either. Criteria of how well ‘a decision theory does on average and over time’, or ‘when a dilemma is issued repeatedly’ run into similar problems as the primary formulation of the criterion. Assessing by how well the decision-maker does in possible worlds that she isn’t in fact in doesn’t seem a compelling criterion (and EDT and CDT could both do well by that criterion, too, depending on which possible worlds one is allowed to pick).

Fourth, arguing that FDT does best in a class of ‘fair’ problems, without being able to define what that class is or why it’s interesting, is a pretty weak argument. And, even if we could define such a class of cases, claiming that FDT ‘appears to be superior’ to EDT and CDT in the classic cases in the literature is simply begging the question: CDT

adherents claims that two-boxing is the right action (which gets you more expected utility!) in Newcomb's problem; EDT adherents claims that smoking is the right action (which gets you more expected utility!) in the smoking lesion. The question is which of these accounts is the right way to understand 'expected utility'; they'll therefore all differ on which of them do better in terms of getting expected utility in these classic cases.

Finally, in a comment on a draft of this note, Abram Demski said that: "The notion of expected utility for which FDT is supposed to do well (at least, according to me) is *expected utility with respect to the prior for the decision problem under consideration*." If that's correct, it's striking that this criterion isn't mentioned in the paper. But it also doesn't seem compelling as a principle by which to evaluate between decision theories, nor does it seem FDT even does well by it. To see both points: suppose I'm choosing between an avocado sandwich and a hummus sandwich, and my prior was that I prefer avocado, but I've since tasted them both and gotten evidence that I prefer hummus. The choice that does best in terms of expected utility with respect to my prior for the decision problem under consideration is the avocado sandwich (and FDT, as I understood it in the paper, would agree). But, uncontroversially, I should choose the hummus sandwich, because I prefer hummus to avocado.

VIII. An alternative approaches that captures the spirit of FDT's aims

Academic decision theorists tends to focus on what actions are rational, but not talk very much about what sort of agent to become. Something that's distinctive and good about the rationalist community's discussion of decision theory is that there's more of an emphasis on what sort of agent to be, and what sorts of rules to follow.

But this is an area where we can eat our cake and have it. There's nothing to stop us assessing agents, acts and anything else we like in terms of our favourite decision theory.

Let's define: *Global expected utility theory* =df for any x that is an evaluative focal point, the right x is that which maximises expected utility.

I think that Global CDT can get everything we want, without the problems that face FDT. Consider, for example, the Prisoner's Dilemma. On the global version of CDT, we can say *both* that (i) the act of defecting is the right action (assuming that the other agent will use their money poorly); and that (ii) the right sort of person to be is one who cooperates in prisoner's dilemmas.

(ii) would be true, even though (i) is true, if you will face repeated prisoner's dilemmas, if whether or not you find yourself in opportunities to cooperate depend on whether or not you've cooperated in the past, if other agents can tell what sort of person you are even independently in your actions in Prisoner's Dilemmas, and so on. Similar things can be said about blackmail cases and about Parfit's Hitchhiker. And similar things can be said more broadly about what sort of person to be given consequentialism — if you become someone who keeps promises, doesn't tell lies, sticks up for their friends (etc), and who doesn't analyse these decisions in consequentialist terms, you'll do more good than someone who tries to apply the consequentialist criterion of rightness for every decision.

(Sometimes behaviour like this is described as 'rational irrationality'. But I don't think that's an accurate description. It's not that one and the same thing (the act) is both

rational and irrational. Instead, we continue to acknowledge that the act is the irrational one; we just also acknowledge that it results from the rational disposition to have.)

There are other possible ways of capturing some of the spirit of FDT, such as a sort of rule-consequentialism, where the right set of rules to follow are those that would produce the best outcome if all agents followed those rules, and the right act is that which conforms to that set of rules. But I think that global causal decision theory is the most promising idea in this space.

IX. Conclusion

In this note, I argued that FDT faces multiple major problems. In my view, these are fatal to FDT in its current form. I think it's possible that, with very major work, a version of FDT could be developed that could overcome some of these problems (in particular, the problems described in sections IV, V and VI, that are based, in one way or another, on the issue of when two processes are Y&S-subjunctively dependent on one another). But it's hard to see what the motivation for doing so is: FDT in any form will violate *Guaranteed Payoffs*, which should be one of the most basic constraints on a decision theory; and if, instead, we want to seriously undertake the project of what decision-procedure is the best for an agent to run (or 'what should we code into an AI?'), the answer will be far messier, and far more dependent on particular facts about the world and the computational resources of the agent in question, than any of EDT, CDT or FDT.

What are the biggest "moonshots" currently in progress?

You know, "moonshot" projects like:

- SpaceX's plan to colonize Mars
- SpaceX's plan to give the whole world fast satellite internet
- Tesla's plan to transition the world to sustainable energy
- Boom making commercial supersonic flight a thing again

Let's make a list!

E.g. does anyone know if China has some huge construction projects that could qualify for this?

Timer Toxicities

Follow-up to: [Free-to-Play Games: Three Key Trade-Offs](#)

The central free-to-play mechanic is to ration action and resources via real world time. This leads to two of the [three key trade offs](#). Players are prevented from having fun because they are time restricted, either unable to play or unable to have the resources to play the way they would like, allowing the game to sell a solution to these problems. More perniciously, players become trained to constantly check in with the game in order to claim rewards and keep their resources from becoming idle. This can warp a person's life more than one would think, changing behavior to allow timely access, and preventing focus on other subjects.

This obsession effect, and the ability of real world time delays to be an interesting resource to include in trade-offs, have also caused these mechanics to seep into non-free games, especially RPGs.

Resource rationing takes the form of timers. The form of the timer does a lot to determine how toxic the rationing will be to the player. There are several knobs one can turn.

The Knobs

Many of these knobs represent related aspects, and thus are closely intertwined, but listing them out still seems useful. In each case, moving towards the first named end will reduce toxicity.

1. Steady versus Sudden: If a resource accumulates over time, does the resource accumulate gradually with no limit, gradually up to a limit, or all at once?
2. Fixed versus Delayed: Does the resource replenish at a fixed time, or at a time after it is used?
3. Slow versus Rapid: How frequently must one check-in to maximize results?
4. Batched versus Disjoint: Are there multiple timers running simultaneously? If so, how hard are they to line up?
5. Tracked versus Lost: How easy is it to track when accumulation is complete?
6. Forgiving versus Punishing: How punishing is it to fail to check in?
7. Isolated versus Cumulative: Are there cumulative rewards for reliably checking in? How important are they?
8. Progressive versus Competitive: Are you in competition with others? If so, what kind?
9. Queued versus Idle: Can you give orders in advance?
10. Explained versus Mysterious: Can you tell what matters?
11. Annoying versus Limiting: Are the things you get from timers the long-term limiting factor in your ability to do things?
12. Incidental versus Central: Is timing the key determinant of your success?

Knob 1: Steady versus Sudden

If you have a resource that replenishes steadily over time, then there is a broad window during which it is efficient to utilize that resource. Even if it has a maximum amount you can store, you can spend it at any point *up to and including* the moment

you have stored up that maximum, with no losses. While you lose potential compound interest gained from whatever the resource might have bought you, and there are usually worries about maximum resource storage, the player can mostly relax most of the time. Even if the resource briefly stops accumulating, this tends to be a relatively light punishment. There is still the worry that spending the resource now is always better than spending it later, as it reduces future risk, and this can line of thinking can still be harmful.

If you have a resource that replenishes all at once, then knob two becomes very important. The risk is that the time of the sudden accumulation becomes something the player worries about missing, and thus schedules around. The player might even sit idle waiting for that time to arrive.

Knob 2: Fixed versus Delayed

If you have a resource that replenishes once per day, at 12:00 midnight Pacific time, you have a full day-long window in which to use that resource. There is no reward or punishment for using it in the morning, afternoon or evening. The player does not feel distracted, and does not have to worry about losing resources when they delay. It does create potential large pressure near the end of the full period, if the resource has yet to be used and cannot be stored.

If you have a resource that replenishes once per day, *twenty-four hours after it is used or claimed*, then every second you do not use the resource is another second it will be delayed, *forever*. If you wait until the time when it is convenient for you to use it, you will need to wait until after that time the next day, and generally run the risk of losing an entire cycle whenever things get delayed at all. There is no stable, relaxed equilibrium available.

If the accumulation is delayed, then that raises the importance of several of the other knobs.

Knob 3: Slow versus Rapid

If full accumulation takes a day, then it is easy to see why this will probably not be too disruptive. Check-in need be at most once per day. A steady *and slow* accumulation is definitely mostly harmless.

If full accumulation takes place every few hours, or less than that, then this can prove extremely disruptive, to the point of costing the player sleep and a constant state of distraction. This is especially true if it is also sudden and the next accumulation is delayed.

The cost of potentially missing a full accumulation can have high emotional resonance. If you are given resources continuously, every moment that this is taken away can be seen as losing out on those resources. If you are going to continue to participate anyway, it can feel extremely bad to let this happen, and seems to be able to trigger the loss aversion circuits in our brains.

Knob 4: Batched versus Disjoint

If your timing triggers are batched, you can respond to each batch as one action.

If your timing triggers are disjoint, you cannot do so without sacrificing efficiency, and the effective tax on your attention and time is much higher.

Suppose your widgets renew every two hours, and your whatsits renew every three hours. You need to check in at hours 2, 3, 4 and 6 every six hour period. You get some benefit of being able to align the second three-hour window with the third two-hour window. If these are delayed triggers, than they might become slightly disjoint, and the intervening time might be wasted literally looking at a countdown timer. The alternative is to not wait around, but if you do that, then the two become fully disjoint.

Consider a game where there are a bunch of different timers and queues – for example, you might have a construction timer, a research timer, an army training timer, a special reward timer and an army task timer. If these always lined up, you might be able to check in (let's say) only three times per day with minimal loss, and each time you'd have a bunch of fun things to do. Instead, you feel bad if you do not check in fifteen times per day, and each time you have only one thing to do.

Juggling does have an upside, *if* you have control over the length of the timers. If you have the ability to choose between different things to build, research or train each of which takes a different amount of time, then you can trade off what you want most per unit time against what will allow you to check back in at the same time? Similar planning problems involving when you sleep, work or are otherwise not going to check in can also be interesting, or they can risk actually disrupting your plans in bad ways. Or both.

Knob 5: Tracked versus Lost

If accumulation is easy to track, that is far less distracting. If you know that accumulation will be sufficient to take action, or will be at maximum levels, or otherwise what you want, at exactly 10:34 AM today, then you can set an alarm, or you can remember to check in at about that time, or you can have your phone notify you, ideally having the game send a notification at that time. The better and easier you can find out and track the right time to do things, the less attention you have to pay to make sure that you'll be there at the right time, so less time and attention is wasted. It is important that games give you this information in forms that are easy to understand and to track.

If accumulation is difficult to track, things can be far worse. An example of this is if the timers are misleading, forcing you to adjust them and to not be able to wait for your notifications. A prominent game I explored allows the player to speed up any timer by spending resources, with the last ten minutes of this process being free. Thus, if something has nine minutes to go, you can click on it to complete it now, so it is essentially finished but for your noticing, and the resource is effectively sitting idle. This ten minute window means that every timer in the game is off, and every notification of this type comes too late, forcing upon the player constant paranoia. It was a relief when, later in the game, average timer length expanded enough that ten minutes was no longer something worth worrying about.

Knob 6: Forgiving versus Punishing

What happens if you fail to check-in at the requested time? Do barbarians burn down your castle? Does your voyage through the stars run out of anti-matter, forcing an abandonment of all rewards or a payment of precious premium currency? Or does

your progression only pause briefly until you are convinced to genuflect in the game's direction in the form of a few clicks?

If you can lose a lot of progress, or even a lot of your resources, by not paying enough attention, that is a very big deal and highly disruptive. You need to have a system where that almost never happens, or the game likely becomes unplayable. If all that happens is your accumulation temporarily halts, the other trade-offs can make this somewhat painful, but it mostly seems fine to play and note care much about full maximization.

It's pretty real-world terrible to hold these types of threats over people's heads. I once read a website that described what machinations one should go through each evening players of a free-to-play game should go through, as non-paying players, *in order to be able to go to sleep at night*.

It is hard to imagine what is being offered in exchange for that, that would make the right response anything but the uninstall button.

Knob 7: Isolated versus Cumulative

If each timer is isolated, then missing one timer and its associated rewards does not impact future timers. Each timer is its own opportunity. The marginal penalty for missing one is small.

If the timers are linked, then the important rewards are usually based on *consistently* checking in for *most or all* of the timers. The marginal benefit of meeting each deadline goes up as you meet more of them, and missing even one often sends you back to square one to start over.

Extreme versions of this stretch rewards over periods of weeks or months, or require intense levels of activity on specified event days or weekends. You have to continuously spend time, then if you don't focus in when they tell you to, most of what you have worked for is lost. The central ideas are the concept of having daily rewards that require not missing timers and then a daily reward for completing all the daily rewards, a login bonus or other reward for doing actions on consecutive days with no misses, and then a 'mastery track' that gives you increasing marginal returns for consistently getting all those daily rewards.

Knob 8: Progressive versus Competitive

If you are free to progress at your own pace in an essentially static world, the game will let you play in a way that is compatible with your life. You do not have to worry that other players will defeat you by making the sacrifices you won't. All that happens if you don't go crazy is that the game gets slower and harder. That might not even be a bad thing.

If you are in a race against time because you face off against other players in a competition to power up the fastest, all starting from roughly the same point as a new shard or region is created, then you are at constant risk of falling behind if you do not maximize your time and money spends, especially if the timers are as central to progression as they usually are in free to play games.

Games that are by their nature competitive races can offer unique and rich experiences. The risk is that when there is not clear separation between the resources

of the game and the resources of your life, where does it end? The central reason the mechanics described in [Meditations on Moloch](#) of sacrificing everything to keep up in zero-sum competitions doesn't actually eat the world – the reason that, to quote an upcoming post that hopefully begins an important sequence, Moloch Isn't Winning, is that such competitions are almost always asymmetrical and complex, with winning not being a direct function of effort and resources, and with the value of winning not being that centrally important.

Games that centrally feature timers are usually *intentionally designed* to topple this barrier. Reward is explicitly locked to effort, in the form of money spent and time devoted, with skill and strategy intentionally crippled in importance beyond a basic level of competence. Thus, all the various complexities that make it not worth sacrificing one's children to Moloch are removed from the equation.

A game's explicit structure of winning and losing also takes away the second most important defense, which is that it's usually not as big a deal to not win such competitions as one might make it out to be. Games are about winning.

Of course, as [Robin Hanson](#) would respond, games are not about winning. Which is also true and important, but doesn't change the true and important, and more relevant in context, fact that games are about winning.

(Leaving a marker here for myself both to eventually fully explain this particular flip, and the general case viewpoint on how to approach when X isn't really about Y but is also of course totally, totally about Y.)

Knob 9: Queued versus Idle

Queues, where you could take action in advance and tell the game what to do when the time came, would mitigate many of these issues. Build queues for towns or buildings in even turn-based strategy games can be the difference between such games being great versus being so tedious as to be almost unplayable. Many mobile games would be much better if, when you tried to do something but lacked the necessary resources, you could click a button that said "do this when resources are available." Other similar tricks could be used in other situations.

Idle however is the universal rule. You need to tell every resource what to do after it becomes available, or it will sit idle until you do.

It's easy to see why idle is chosen over queues. The whole point of the system, as we'll discuss after the list is finished, is to force frequent check-in and obsession to facilitate habit formation, addiction and obsession. This potential solution is solving the player's problem, not the game's problem.

Knob 10: Explained versus Mysterious

This is a variation of Tracked versus Lost. If you don't even know when checking in will be rewarded, then you certainly can't track it. With Lost, you have a hard time remembering or recording exactly when something will happen, but you likely know more or less what will happen and about when it will happen. If things are sufficiently mysterious, you can feel an obligation to continuously check *in case* something happens, without any theory as to what or when or how. Eventually it becomes a compulsion without a justification.

Knob 11: Annoying versus Limiting

Knob 12: Incidental versus Central

These last two knobs ask how much the game is fundamentally about its timers.

Annoying timers mean that your life will be worse off if you miss them or mismanage them, but *over the long term* the number of check-ins is not the determining limiting factor of progress. You'll end up in the same place. For limiting, that means that the timers give you the main source of the resource or resources that are the limiting factor for your progression. If there are severely limiting factors, it is quite possible for most other resources and accomplishments to not matter.

Incidental versus central compares the generic value of what the timers give you to the generic value of what is otherwise available. In the extreme case, the key limiting resource is *only* available via timers, or even timers with checking in as the only requirement, and the rest of the game does not actually exist. Instead you end up in the situation described in a comment on my previous post, by Villam:

I will not mention the name of the game here. Anyway, it was the type of game where you build stuff, collect resources, and research new stuff; with many things to unlock. In the game there were three important resources, let's call them X, Y, and Z. By making better or worse decisions, you could make more or less of the resources X and Y; and I spent some time optimizing for that.

With resource Z, however, the basic way to get it was to *play the game regularly*. If you logged in at least N times a day, you got M points of resource Z per day; you couldn't get more for playing longer, but you would get less for taking breaks longer than $1/N$ of the day. In addition to this, there were also some other ways to get resource Z, but this extra amount was always *smaller* than the amount you got for merely playing the game regularly. There was no smart strategy to at least double the income of Z. So, whether you did smart or stupid things had a visible impact on X and Y, but almost *no impact* on Z.

Of course the resource Z was the one that actually mattered, in long term. Your progress on the tech tree sometimes required X and Y, but *always* required Z. And, of course, the higher steps on the almost-linear tech tree required *more* of the resource Z.

A popular variation of this is to make your progress impact the rate at which Z is collected when you check in. It then matters how efficiently you navigate the early stages in order to level yourself up, because it increases your timer rewards, and that is the primary thing to optimize for.

Of course, all of that presupposes that your goal is to make your numbers go up as high as possible *over a very long term of calendar time* in a world in which the purpose of numbers going up is... to make them then go slightly higher than that, at ever slower rates. Not exactly the most enticing proposition, when stated that way.

Ask this question: Would you do better long term if you did nothing today but check-in for the timers, or if you did everything else but didn't check-in with the timers today?

As The Knob Turns: Toxicity versus Compulsion

Games with timers are using real world time as a managed resource. As noted, this can be an interesting design space and set of optimization problems.

I see three categories of toxicity trade-offs regarding timers.

There are timers that require toxicity because they are putting real world attention into the game as a resource constraint.

There are games that are not trying to do that but which have to pay costs to avoid imposing costs.

Then there are most of the games in the genre, which are mostly using Skinner box tactics to create habits and compulsive behaviors. Toxicity is not a bug, *it is a feature*. It is the hill that they climb. It is the killer app. Their goal is to turn all twelve knobs as far to the right as possible without players taking too much notice.

The first case is sympathetic. I do believe that it is real and legitimate, and not merely a cover for the third case, but its presence in its good form is rare. An interesting choice requires trade-offs, so the desire to minimize toxicity and real world cost in exchange for in-game benefits becomes the real game. Done right, that's cool, and you can't have those real world stakes without at least some real world costs. There is a real trade-off. My best advice for a designer in these situations is to ensure that there are always reasonable solutions that impose only reasonable real world costs, and benefits of further attention have to decline rapidly beyond roughly the commit the player is intentionally making. It is also important that optimization *works*, with better approaches being faster and requiring less attention and check-in than poor approaches.

The second case is also sympathetic, and in its full form is common. In addition to real world clock time being an interesting resource to trade-off, letting players sample the game at some rate, while imposing costs for moving faster, is a relatively friendly business model. You would prefer to avoid distracting the player outside of time intentionally dedicated to your game, and you would prefer to have that dedicated time focused on interesting game play decisions rather than engaging in micromanagement. Queues are helpful here, as are many of the other knobs, which can safely be turned far to the left.

The third case is, unfortunately, the central case. Toxicity is turned on its head and embraced as the core game feature.

From the players' perspective, one must figure out how to navigate these dangers, and whether there is a path to doing so without the game becoming net negative. When deciding this, one needs to keep in mind that these systems are designed to hijack your brain, developing habits and compulsions that may be difficult to break. Thus, if things turn out to be bad, that badness is designed to prevent you from realizing this or from being able to execute on disengagement. One cannot do the calculation assuming one can think clearly in the future.

Last time, I recommended looking at how often you were required to ping a game, and how punishing it was to fail to do so, as a way to estimate a game's toxicity level. This remains a good simple heuristic. Looking at the knobs above allows you to flesh out this evaluation. When looking at the games I have been surveying, this made me realize that I had been fooled by at least one game into thinking it was far less toxic and far less relatively toxic than it was, and that I had to adjust modes of further play to reduce its toxicity level.

From the designer's perspective, whether or not you have ethical concerns, the question is how to balance the costs you impose on the player, the negative reactions the player will have to those costs, and the upside from getting players addicted and in the habit of playing your game every day and in any spare moment. Here is where we see that in each of the twelve cases, turning the knob to the right increases the rate at which habit and compulsion are imposed upon the player. Thus, the limiting factor will by default be what players will accept without running away screaming or otherwise realizing the game is not their friend.

This leads into the question of how you measure that. I have been thinking a lot about what happens when you have very strong ability to measure specific short term outcomes, but much less ability to measure other longer term outcomes, and there are implicit hidden variables.

Consider the following toy model of a free-to-play game player.

The player operates based on hidden variables. They have ongoing levels of (related) things like Fun, Compulsion, Habit, Willingness to Pay, Willingness to Recommend, Annoyance, Goodwill, Trust, Social Ties, Sense of Accomplishment, and so on.

The game cannot measure these directly. Instead, the game gets to measure things like hours played, what causes players to log out or keep playing in the moment, when players spend money, when players uninstall, how players manage their resources, review star ratings, and so on.

Any given decision will get made largely based on optimizing those short term outcomes. Short term outcomes, especially time played and money spent, will loom large. It will be tough to justify sacrificing those outcomes to lay longer term foundations that may or may not exist or matter, or might even go in the opposite direction.

This is especially true because the correlations are strong between the observed short term outcomes and expected long term outcomes. There are even strong causal mechanisms explaining why good observed results now lead directly to good results later. Players who spend more now will spend more later. Players who play more now form the habits and compulsions to spend more later, as they 'break the seal' and identify themselves as people who spend money in this way. Interacting with friends or joining a guild leads to more social engagement.

Combine that with the ease of optimizing between known options via A/B testing in order to hill climb on the details that lead to superior metrics, and you know that the special limited time offer will be the exact correct size and color and price and duration and timing and so on that maximizes short term profits.

Measuring the formation of habits and compulsions is harder than figuring out what color causes more sales in the next hour, but it still offers solid metrics. You should see the trend lines moving in the appropriate ways almost right away, or the effect is not there. The downside costs all this imposes can be divided into two categories. There are the times that you flip a switch in someone's head that says 'this game is super toxic and I need to run away screaming right now.' You learn to avoid doing that. However, there is also the gradual accumulation of annoyances and loss of trust and goodwill that comes from not giving players a good experience, or from imposing steady hard to notice costs upon them, or by being too repetitive, or what not. This sneaks up on you slowly, and any given decision is usually going to have a small impact that is not going to be directly measurable, but that will all add up over time.

These and other related Goodhart's Law problems explain a lot of what I see as clear failures to optimize player experience, and experience in other services and places across the internet and all of our civilization, in ways I plan to explore more, but I will stop here to avoid too much scope creep.

The biggest question one wants to ask is, to what extent are you in case one versus case two versus case three. If you are in the first two cases, toxicity can be better or worse, but you should not expect toxicity levels to continue to rise, or for them to be especially deceptively high. If you are in the third case, you should assume that things are already worse than you think and designed to get steadily worse than that.

Thus, one should look for knobs that are *intentionally turned to the right*. If there are mechanics that seem designed to force additional check-in times, to punish failure to check-in, to force you to form reliable habits or lose out on the bulk of rewards, and so on, that should be a big red flag. Combining this with looking at expected rates of check-in should give a good picture. Going down the line on all the knobs should give a better picture still.

The next step in this exploration is to look at particular game mechanics in detail.

Who's an unusual thinker that you recommend following?

Some (optional!) desiderata to guide your recommendation:

- They *probably* have not been heard of around these parts.
- They disagree with a consensus.
- They have genuinely original thoughts.
- You have updated from their views and/or been emotionally affected.
- They're actively doing some interesting thing(s) and seem effective at those things.
- Reading/listening doesn't feel like an obtuse poetic maze; they attempt to convey things relatively clearly.
- Maybe they use an unusual method of communication.

Feel free to *disregard* these criteria if it doesn't work for answering the question! :)

Idols of the Mind Pt. 2 (Novum Organum Book 1: 53-68)

This is the fifth post in the [Novum Organum sequence](#). For context, see [the sequence introduction](#).

We have used Francis Bacon's *Novum Organum* in the version presented at www.earlymoderntexts.com. Translated by and copyright to [Jonathan Bennett](#). Prepared for LessWrong by [Ruby](#).

Ruby's Reading Guide

Novum Organum is organized as two books each containing numbered "aphorisms." These vary in length from three lines to sixteen pages. Titles of posts in this sequence, e.g. *Idols of the Mind Pt. 1*, are my own and do not appear in the original.

While the translator, Bennett, encloses his editorial remarks in a single pair of [brackets], I have enclosed mine in a [[double pair of brackets]].

Bennett's Reading Guide

[Brackets] enclose editorial explanations. Small ·dots· enclose material that has been added, but can be read as though it were part of the original text. Occasional •bullets, and also indenting of passages that are not quotations, are meant as aids to grasping the structure of a sentence or a thought. Every four-point ellipsis indicates the omission of a brief passage that seems to present more difficulty than it is worth. Longer omissions are reported between brackets in normal-sized type.

Aphorism Concerning the Interpretation of Nature: Book 1: 53-68

by Francis Bacon

53. The idols of the cave—my topic until the end of **58**—arise from the particular mental and physical make-up of the individual person, and also from upbringing, habits, and chance events. There are very many of these, of many different kinds; but I shall discuss only the ones we most need to be warned against—the ones that do most to disturb the clearness of the intellect.

54. A man will become attached to one particular science and field of investigation either because •he thinks he was its author and inventor or because •he has worked hard on it and become habituated to it. But when someone of this kind turns to *general* topics in philosophy ·and science· he wrecks them by bringing in distortions from his former fancies. This is especially visible in Aristotle, who made his natural science a mere bond-servant to his logic, rendering it contentious and nearly useless. The chemists have taken a few experiments with a furnace and made a fantastic science out of it, one that applies to hardly anything. . . .

[In this work 'chemists' are alchemists. Nothing that we would recognize as chemistry existed.]

[[We might see Bacon here as claiming that "seeing everything as a nail" can be very harmful.]]

55. When it comes to philosophy and the sciences, minds differ from one another in one principal and fairly radical way: some minds have more liking for and skill in •noting differences amongst things, others are adapted rather to •noting things' resemblances. The •steady and acute mind can concentrate its thought, fixing on and sticking to the subtlest distinctions; the •lofty and discursive mind recognizes and puts together the thinnest and most general resemblances. But each kind easily goes too far: one by •grasping for ·unimportant· differences between things, the other by •snatching at shadows.

56. Some minds are given to an extreme admiration of antiquity, others to an extreme love and appetite for novelty. Not many have the temperament to steer a middle course, not pulling down sound work by the ancients and not despising good contributions by the moderns. The sciences and philosophy have suffered greatly from this, because these attitudes to antiquity and modernity are not *judgments* but mere *enthusiasms*. Truth is to be sought not in •what people like or enjoy in this or that age, but in •the light of nature and experience. The •former is variable, the •latter is eternal. So we should reject these enthusiasms, and take care that our intellect isn't dragged into them.

57. When you think ·hard and long and uninterrupted· about nature and about bodies in their simplicity—i.e. think of topics like *matter as such*—your intellect will be broken up and will fall to pieces. When on the other hand you think ·in the same way· about nature and bodies in all their complexity of structure, your intellect will be

stunned and scattered. The difference between the two is best seen by comparing the school of Leucippus and Democritus with other philosophies. For the members of that school were so busy with the ·general theory of· particles that they hardly attended to the structure, while the others were so lost in admiration of the structure that they didn't get through to the simplicity of nature. What we should do, therefore, is alternate between these two kinds of thinking, so that the intellect can become *both* penetrating *and* comprehensive, avoiding the disadvantages that I have mentioned, and the idols they lead to.

58. Let that kind of procedure be our prudent way of keeping off and dislodging the idols of the cave, which mostly come from

- intellectual· favouritism (**54**),
- an excessive tendency to compare or to distinguish (**55**),
- partiality for particular historical periods (**56**), or
- the largeness or smallness of the objects contemplated (**57**).

Let every student of nature take this as a general rule for helping him to keep his intellect balanced and clear: when your mind seizes on and lingers on something with special satisfaction, treat it with suspicion!

59. The idols of the market place are the most troublesome of all—idols that have crept into the intellect out of the contract concerning words and names [Latin *verborum et nominum*, which could mean ‘verbs and nouns’; on the contract, see **43**]. Men think that their reason governs words; but it is also true that words have a power of their own that reacts back onto the intellect; and this has rendered philosophy and the sciences sophistical and idle. Because words are usually adapted to the abilities of the vulgar, they follow the lines of division that are most obvious to the vulgar intellect. When a language-drawn line is one that a sharper thinker or more careful observer would want to relocate so that it suited the true divisions of nature, words stand in the way of the change. That's why it happens that when learned men engage in high and formal discussions they often end up arguing about words and names, using definitions to sort them out—thus •ending where, according to mathematical wisdom and mathematical practice, it would have been better to •start! But when it comes to dealing with natural and material things, definitions can't cure this trouble, because the definitions themselves consist of words, and those words beget others. So one has to have recourse to individual instances. . . .

[[Bacon grokked that misuses of words were a great cause of confusion. He probably would have like the [A Human's Guide to Words Sequence](#). See [Where to Draw the Boundary?](#) and [37 Ways That Words Can Be Wrong.](#)]]

60. The idols that words impose on the intellect are of two kinds. (1) There are names of things that don't exist. Just as there are things with no names (because they haven't been observed), so also there are names with no things to which they refer—these being upshots of fantastic ·theoretical· suppositions. Examples of names that owe their origin to false and idle theories are ‘fortune’, ‘prime mover’, ‘planetary orbits’, and ‘element of fire’. This class of idols is fairly easily expelled, because you can wipe them out by steadily rejecting and dismissing as obsolete all the theories ·that beget them·.

[[See [Empty Labels](#).]]

(2) Then there are names which, though they refer to things that do exist, are confused and ill-defined, having been rashly and incompetently derived from realities.

Troubles of this kind, coming from defective and clumsy abstraction, are intricate and deeply rooted. Take the word ‘wet’, for example. If we look to see how far the various things that are called ‘wet’ resemble one other, we’ll find that ‘wet’ is nothing but than a mark loosely and confusedly used to label a variety of states of affairs that can’t be unified through any constant meaning. For something may be called ‘wet’ because it

- easily spreads itself around any other body,
- has no boundaries and can’t be made to stand still,
- readily yields in every direction.
- easily divides and scatters itself,
- easily unites and collects itself,
- readily flows and is put in motion,
- readily clings to another body and soaks it,
- is easily reduced to a liquid, or (if it is solid) easily melts.

Accordingly, when you come to apply the word, if you take it in one sense, flame is wet; if in another, air is not wet; if in another, fine dust is wet; if in another, glass is wet. So that it is easy to see that the notion has been taken by abstraction only from water and common and ordinary liquids, without proper precautions.

Words may differ in *how* distorted and wrong they are. One of the •least faulty kinds is that of names of substances, especially names that

- are names of lowest species, ·i.e. species that don’t divide into sub-species·, and
- have been *well* drawn ·from the substances that they are names of·.

·The drawing of substance-names and -notions from the substances themselves *can* be done well or badly. For example·, our notions of chalk and of mud are good, our notion of earth bad. •More faulty are names of events: ‘generate’, ‘corrupt’, ‘alter’. •The most faulty are names of qualities: ‘heavy’, ‘light’, ‘rare’, ‘dense’, and the like. (I exclude from this condemnation names of qualities that are immediate objects of the senses.) Yet in each of these categories, inevitably some notions are a little better than others because more examples of them come within range of the human senses.

61. The idols of the theatre ·which will be my topic until the end of **68**· are not innate, and they don’t steal surreptitiously into the intellect. Coming from the fanciful stories told by philosophical theories and from upside-down perverted rules of demonstration, they are openly proclaimed and openly accepted. Things I have already said imply that there can be no question of *refuting* these idols: where there is no agreement on premises or on rules of demonstration, there is no place for argument.

·AN ASIDE ON THE HONOUR OF THE ANCIENTS·

This at least has the advantage that it leaves the honour of the ancients untouched ·because I shall not be *arguing against* them. I shall be *opposing* them, but· there will be no disparagement of them in this, because the question at issue between them and me concerns only *the way*. As the saying goes: a lame man on the right road outstrips the runner who takes a wrong one. Indeed, it is obvious that a man on the wrong road goes further astray the faster he runs. ·You might think that in claiming to be able to do better in the sciences than they did, I must in some way be setting myself up as brighter than they are; but it is not so·. The course I propose for discovery in the sciences leaves little to the acuteness and strength of intelligence, but puts all intelligences nearly on a level. My plan is exactly like the drawing of a straight line or a perfect circle: to do it free-hand you need a hand that is steady and

practised, but if you use a ruler or a compass you will need little if anything else; and my method is just like that.

·END OF ASIDE·

But though particular counter-arguments would be useless, I should say something about •the classification of the sects whose theories produce these idols, about •the external signs that there is something wrong with them, and lastly •about the causes of this unhappy situation, this lasting and general agreement in error. My hope is that this will make the truth more accessible, and make the human intellect more willing to be cleansed and to dismiss its idols.

62. There are many idols of the theatre, or idols of theories, and there can be and perhaps will be many more. For a long time now two factors have militated against the formation of new theories ·in philosophy and science·.

- Men's minds have been busied with religion and theology.
- Civil governments, especially monarchies, have been hostile to anything new, even in theoretical matters; so that men have done that sort of work at their own peril and at great financial cost to themselves—not only unrewarded but exposed to contempt and envy.

If it weren't for those two factors, there would no doubt have arisen many other philosophical sects like those that once flourished in such variety among the Greeks. Just as many hypotheses can be constructed regarding the phenomena of the heavens, so also—and even more!—a variety of dogmas about the phenomena of philosophy may be set up and dug in. And something we already know about plays that poets put on the stage is also true of stories presented on the philosophical stage —namely that fictions invented for the stage are more compact and elegant and generally liked than true stories out of history!

What has gone wrong in philosophy is that it has attended in great detail to a few things, or skimpily to a great many things; either way, it is based on too narrow a foundation of experiment and natural history, and decides on the authority of too few cases. **(1)** Philosophers of the reasoning school snatch up from experience a variety of common kinds of event, without making sure they are getting them right and without carefully examining and weighing them; and then they let meditation and brain-work do all the rest. **(2)** Another class of philosophers have carefully and accurately studied a few experiments, and have then boldly drawn whole philosophies from *them*, making all other facts fit in by wildly contorting them. **(3)** Yet a third class consists of those who are led by their faith and veneration to mix their philosophy with theology and stuff handed down across the centuries. Some of these have been so foolish and empty-headed as to have wandered off looking for knowledge among spirits and ghosts. So there are the triplets born of error and false philosophy: philosophies that are **(1)** sophistical, **(2)** empirical, and **(3)** superstitious.

[To explain Bacon's second accusation against Aristotle in **63**: A word 'of the second intention' is a word that applies to items of thought or of language (whereas things that are out there in the world independently of us are referred to by words 'of the first intention'). Now Aristotle in his prime held that the soul is not a *substance* but rather a *form*: rather than being an independently existing thing that is somehow combined with the rest of what makes up the man, the soul is a set of facts about how the man acts, moves, responds, and so on. Bacon has little respect for the term 'form': in **15** he includes it among terms that are 'fantastical and ill-defined', and in **51** he

says that ‘forms are fabrications of the human mind’. This disrespect seems to underlie the second accusation; the class of *forms* is not a class of independently existing *things* but rather a class of muddy and unfounded *ways of thinking and talking*, so that ‘form’ is a word of the second intention.]

63. The most conspicuous example of (1) the first class was Aristotle, whose argumentative methods spoiled natural philosophy. He

- made the world out of categories;
- put the human soul, the noblest of substances, into a class based on words of the second intention;
- handled the issues about density and rarity (which have to do with how much space a body takes up) in terms of the feeble distinction between what does happen and what could happen;
- said that each individual body has one proper motion, and that if it moves in any other way this must be the result of an external cause,

and imposed countless other arbitrary restrictions on the nature of things. He was always less concerned about the inner truth of things than he was about providing answers to questions—*saying* something definite. This shows up best when his philosophy is compared with other systems that were famous among the Greeks. For

- the homogeneous substances of Anaxagoras,
- the atoms of Leucippus and Democritus,
- the heaven and earth of Parmenides,
- the strife and friendship of Empedocles, and
- Heraclitus’s doctrine of bodies’ being reduced to the perfectly homogeneous condition of fire and then remolded into solids,

all have a touch of natural philosophy about them—a tang of the nature of things and experience and bodies. Whereas in Aristotle’s physics you hear hardly anything but the sounds of logical argument—involving logical ideas that he reworked, in a realist rather than a nominalist manner, under the imposing name of ‘metaphysics’. Don’t be swayed by his frequent mentions of experiments in his *On Animals*, his *Problems*, and others of his treatises. For he didn’t consult experience, as he should have done, *on the way to his decisions and first principles*; rather, he first *decided* what his position would be, and *then brought* in experience, twisting it to fit his views and making it captive. So on this count Aristotle is even more to blame than his modern followers, the scholastics, who have abandoned experience altogether.

64. The (2) empirical school of philosophy gives birth to dogmas that are more deformed and monstrous than those of the sophistical or reasoning school. The latter has as its basis the •light of vulgar notions; it’s a faint and superficial light, but it is in a way •universal, and applies to many things. In contrast with that, the empirical school has its foundation in the •narrowness and •darkness of a few experiments. Those who busy themselves with these experiments, and have infected their imagination with them, find such a philosophy to be probable and all but certain; everyone else finds them flimsy and incredible. A notable example of this •foolishness• is provided by the alchemists and their dogmas; these days there isn’t much of it anywhere else, except perhaps in the philosophy of Gilbert. Still, I should offer a warning relating to philosophies of this kind. If my advice ever rouses men to take experiments seriously and to bid farewell to sophistical doctrines, then I’m afraid that they may—I foresee that they *will*—be in too much of a hurry, will leap or fly •from experiments straight• to generalizations and principles of things, risking falling into

just the kind of philosophy I have been talking about. We ought to prepare ourselves against this evil now, ·well in advance·.

65. The corruption of philosophy by **(3)** superstition and input from theology is far more widespread, and does the greatest harm, whether to entire systems or to parts of them. ·Systems thus afflicted are just nonsense judged by ordinary vulgar standards, but that doesn't protect men from accepting them, because· the human intellect is open to influence from the imagination as much as from vulgar notions, ·and in these philosophies it is the imagination that wields the power·. Whereas the contentious and sophistical kind of philosophy combatively *traps* the intellect, this ·superstitious· kind, being imaginative and high-flown and half-poetic, *coaxes* it along. For men—especially intelligent and high-minded ones—have intellectual ambitions as well as ambition of the will.

A striking example of this sort of thing among the Greeks is provided by Pythagoras, though ·his form of it wasn't so dangerous, because· the superstition that he brought into it was coarser and more cumbersome ·than many·. Another example is provided by Plato and his school, whose superstition is subtler and more dangerous. Superstition turns up also in parts of other philosophies, when they

- introduce abstract forms—·i.e. forms that aren't the forms of anything·,

and when they do things like

- speaking of 'first causes' and 'final causes' and usually omitting *middle* causes.

[Bacon's point is: They discuss the first cause of the whole universe, and the end or purpose for which something happens (its 'final cause'), but they mostly ignore ordinary *causes* such as spark's causing a fire. Putting this in terms of first-middle-final seems to be a quiet joke].

We should be *extremely* cautious about this. There's nothing worse than the *deification* of error, and it is a downright plague of the intellect when empty nonsense is treated with veneration. Yet some of the moderns have been so tolerant of this emptiness that they have—what a shallow performance!—tried to base a system of natural philosophy on the first chapter of Genesis, on the book of Job, and other parts of the sacred writings, 'seeking the living among the dead' [Luke 24:5]. This makes it more important than ever to keep down this ·kind of philosophy·, because this unhealthy mixture of human and divine gives rise not only to ·fantastic philosophy but also to ·heretical religion. It is very proper that we soberly give our faith only to things that are the faith.

66. So much for the mischievous authority of systems founded on ·vulgar notions, on ·a few experiments, or on ·superstition. I should say something about bad choices of what to think *about*, especially in natural philosophy. In the mechanical arts the main way in which bodies are altered is by composition or separation; the human intellect sees this and is infected by it, thinking that something like it produces all alteration in the universe. This gave rise to ·the fiction of *elements* and of their coming together to form natural bodies. Another example: When a man surveys nature working freely, he encounters different species of things—of animals, of plants, of minerals—and that leads him smoothly on to the opinion that nature contains certain *primary forms* which nature intends to work with, and that all other variety comes from ·nature's being blocked and side-tracked in her work, or from ·conflicts between different species—conflicts in which one species turns into another. To the first of these theories we owe ·such intellectual rubbish as· *first qualities of the elements*; to the second we owe

occult properties and *specific virtues*. Both of them are empty short-cuts, ways for the mind to come to rest and not be bothered with more solid pursuits. The medical researchers have achieved more through their work on the second qualities of matter, and the operations of attracting, repelling, thinning, thickening, expanding, contracting, scattering, ripening and the like; and they would have made much greater progress still if *it weren't for a disaster that occurred. The two short-cuts that I have mentioned (elementary qualities and specific virtues) snared the medical researchers, and spoiled what they did with their correct observations in their own field.

[The passage flagged by asterisks expands what Bacon wrote, in ways that the small-dots system can't easily indicate.]

It led them either •to treating second qualities as coming from highly complex and subtle mixture of first or elementary qualities, or •to breaking off their empirical work prematurely, not following up their observations of second qualities with greater and more diligent observations of third and fourth qualities.* ·This is a bigger disaster than you might think, because· something like—I don't say exactly like—the powers involved in the self-healing of the human body should be looked for also in the changes of all other bodies.

But something much worse than that went wrong in their work: they focused on

- the principles governing things at rest, not on •the principles of change; i.e. on
- what things are produced *from*, not •*how* they are produced; i.e. on
- topics that they could talk about, not •ones that would lead to results.

The vulgar classification of ·kinds of· motion that we find in the accepted system of natural philosophy is no good—I mean the classification into

- generation,
- corruption,
- growth,
- diminution,
- alteration, and
- motion.

Here is what they mean. If a body is moved from one place to another without changing in any other way, this is •motion; if a body changes qualitatively while continuing to belong to the same species and not changing its place, this is •alteration; if a change occurs through which the mass and quantity of the body don't remain the same, this is •growth or •diminution; if a body is changed so much that it changes substantially and comes to belong to a different species, this is •generation or •corruption. But all this is merely layman's stuff, which doesn't go at all deeply into nature; for these are only *measures* of motion. . . .and not *kinds* of motion. They [= the notions involved in the classification into generation, corruption etc.] signify that the motion went this way or that, but not *how* it happened or what *caused* it. They tell us nothing about the appetites of bodies [= 'what bodies are naturally disposed to do'] or about what their parts are up to. They come into play only when the motion in question makes the thing grossly and obviously different from how it was. Even when ·scientists who rely on the above classificatory system· do want to indicate something concerning the *causes* of motion, and to classify motions on that basis, they *very* lazily bring in the ·Aristotelian· distinction between 'natural' motion and 'violent' motion, a distinction that comes entirely from vulgar ways of thinking. In fact, 'violent'

motion is natural motion that is called ‘violent’ because it involves an external cause working (naturally!) in a different way from how it was working previously.

[Bacon himself sometimes describes a movement as *violens*, but this is meant quite casually and not as a concept belonging to basic physics. These innocent occurrences of *violens* will be translated as ‘forceful’.]

Let us set all this aside, and consider such observations as that bodies have an appetite for

mutual contact, so that separations can’t occur that would break up the unity of nature and allow a vacuum to be made;

or for

resuming their natural dimensions. . . . , so that if they are compressed within or extended beyond those limits they immediately try to recover themselves and regain their previous size;

or for

gathering together with masses of their own kind—e.g. dense bodies moving towards the earth, and light and rare bodies towards the dome of the sky.

These and their like are truly *physical* kinds of motion; and comparison of them with the others that I mentioned makes clear that the others are entirely *logical* and *scholastic*.

An equally bad feature of their philosophies and their ways of thinking is that all their work goes into investigating and theorizing about the

- fundamental principles of things. . . .—so they keep moving through higher and higher levels of abstraction until they come to *formless potential matter*—and
- the ultimate parts of nature—so they keep cutting up nature more and more finely until they come to *atoms*, which are too small to contribute anything to human welfare—

whereas everything that is useful, everything that can be worked with, lies between those two extremes.

67. The intellect should be warned against the intemperate way in which systems of philosophy deal with the giving or withholding of assent, because intemperance of this kind seems to establish idols and somehow prolong their life, leaving no way open to reach and dislodge them.

There are two kinds of excess: •the excess of those who are quick to come to conclusions, and make sciences dogmatic and lordly; and •the excess of those who deny that we can know anything, and so lead us into an endlessly *wandering* kind of research. The •former of these subdues the intellect, the •latter deprives it of energy. The philosophy of Aristotle is of the former kind. Having destroyed all the other philosophies in argumentative battle. . . . Aristotle laid down the law about everything, and then proceeded to raise new questions of his own and to dispose of them likewise, so that everything would be certain and settled—a way of going about things that his followers still respect and practice.

The ·Old Academy·, the school of Plato, introduced acatalepsy—the doctrine that nothing is capable of being understood. At first it was meant as an ironical joke at the expense of the older sophists—Protagoras, Hippias, and the rest—whose greatest fear was to seem *not to doubt* something! But the New Academy made a dogma of acatalepsy, holding it as official doctrine. They did allow of some things to be followed as probable, though not to be accepted as true; and they said they didn't ·mean to· destroy all investigation; so their attitude was better than. . . .that of Pyrrho and his sceptics. (It was also better than undue freedom in making pronouncements.) Still, once the human mind has despaired of finding truth, it becomes less interested in everything; with the result that men are side-tracked into pleasant disputationes and discourses, into *roaming*, rather than severely sticking to a single course of inquiry. But, as I said at the start and continue to urge, the human senses and intellect, weak as they are, should not be •deprived of their authority but •given help.

68. So much for the separate classes of idols and their trappings. We should solemnly and firmly resolve to deny and reject them all, cleansing our intellect by freeing it from them. Entering the kingdom of man, which is based on the sciences, is like entering the kingdom of heaven, which one can enter only as a little child.

Edited: The next post in the sequence, Book 1: 69-92 (13 Causes of Bad Science), will be posted Thursday, October 3rd at latest by 6:00pm PDT.

Concrete experiments in inner alignment

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This post is part of research I did at OpenAI with mentoring and guidance from Paul Christiano.

The goal of this post is to present my thoughts on some of the sorts of experiments that might be able to be done now that could shed light on the [inner alignment problem](#). I've been doing a lot of thinking about inner alignment from a theoretical perspective recently that I'm pretty excited about, but at the same time, I think there's also a lot of concrete experimental work that can be done in this space as well. That being said, this post is mostly just a brain dump; I expect a lot of additional work will have to be put in to actually take any of these proposals across the finish line.

If you're interested in working on any of these proposals, however, feel free to just go ahead and take it on—you don't need my permission to do so!^[1] That being said, if you'd like to talk to me about one of them—which I would love to do if you're thinking of seriously working on one of these ideas—please do reach out to me either in the comments here or by sending me an email at evanjhub@gmail.com.

Concrete proposals

Reward side-channels

Proposal: Train an RL agent with access to its previous step reward as part of its observation. Then, at test time, modify the observed reward. Measure to what extent the agent continues optimizing the original reward versus switches to optimizing the new observed reward. Compare to the situation where the agent is not given the reward at all and the situation where the agent is given the next step reward (for each action in a discrete action space) rather than the previous step reward. As an implementation detail, I think the RL agent should include either LSTMs or Transformers to enable it to keep track of its multi-step return if it wants to.

Motivation: One important question for understanding the likelihood of deceptive or corrigible alignment vs. robust alignment is to what extent models tend to learn their goals internally vs. via reference to things in their environment. This experiment directly attacks that question by asking whether/to what extent an RL agent will learn to optimize a reward signal in its environment. This is relevant both for understanding how to train corrigibility as well as how to avoid deceptive alignment.

Extensions: Add noise to the observed reward signal and/or try replacing the observed reward signal with some function of the reward instead such as a randomly initialized neural network.

Cross-episodic objectives

Proposal: Train an RL agent in an environment that contains a side-channel for boosting the agent's reward in the next episode. Measure to what extent the agent takes advantage of it. Try different population-based training approaches. Measure again.

Motivation: Safety approaches like amplification and debate depend heavily on agents doing solely myopic optimization, whereas approaches like IRL and reward modeling depend on long-term forward-looking optimization. Thus, understanding the conditions under which agents exploit non-myopic reward side channels could be very useful for shedding some light on the best training techniques to use for different alignment approaches.

Extensions: Instead of getting a cross-episodic objective via population-based training, try just giving the model information that makes it obvious that this is possible such that even a naive planning algorithm would be able to deduce that it can get extra reward across episodes this way. As an implementation detail, I think you will need to use some sort of optimization-as-a-layer to make this work.

Objective unidentifiability

Proposal: Train an RL agent in an environment with multiple simple, obvious objectives all of which would equally explain the true reward. Test in environments that distinguish between the different possible objectives and determine the situations under which models tend towards some objectives versus others. In particular, look for situations in which the model learns some proxy which it pursues competently off-distribution but which has poor off-distribution performance on the true reward. As an implementation detail, I expect you might need to use some sort of optimization-as-a-layer for this to work.

Motivation: Knowing what sorts of proxies models tend to use in different situations would enable us to have a better understanding of the most likely forms of pseudo-alignment and thus the best ways to push models towards robust alignment instead. Furthermore, being able to clearly demonstrate proxy alignment could help in making the case for inner alignment being a real issue.

Extensions: Try changing inductive biases and/or model capacity and see if it changes the types of proxies that the model uses.

Zero-shot objectives

Proposal: Set up a system that enables a language model to take actions in an environment to optimize some reward. Do IRL on the resulting behavior and inspect the objective that results. Find some way to measure how coherent it is. Compare to an RL agent trained directly on the reward. As an implementation detail, you should focus on environments where a language model might actually be able to perform well. (Edit: I think my explanation here was a bit too general, so here's an example of what I mean. Take a language model and see if you can get it to perform tasks which traditionally require RL/search/optimization and see if it can do them. For example, can GPT-2 play Chess? What about a new game not in the training data that you give it the rules to? If these sorts of things work, then that seems to imply that your model must have learned some sort of optimization algorithm, if only some sort of minimax algorithm for searching for the best moves in a game.)

Motivation: It is possible that the best model that maximizes predictive accuracy in a language context might be—at least to some extent—goal-directed. Though current language models are likely not advanced enough to show true goal-directed behavior, it might at least be possible to see the beginnings of coherent objectives.

Extensions: Experiment with different ways to encode RL observations and actions as text and see how consistent the resulting IRL objectives are.

Robust reward learning

Proposal: Train a reward-predictive model-based RL agent (e.g. an [Imagination-based Planner](#)) and compare the resulting objective to the actual reward. Train again with adversarial training on inputs that produce maximally differing reward estimates and compare again.

Motivation: Useful for testing the ability of adversarial training to resolve reward unidentifiability as well as providing information on to what extent adversarial training for aligning the reward model produces aligned actions.

Extensions: Try relaxing the adversary to output pseudo-inputs (e.g. activations) that would lead to maximally differing assessments instead of inputs and backpropagate on those instead.

Other thoughts

This section is just a list of all the different possible testable hypotheses I came up with when thinking about this—including those that made it into the above list of concrete proposals and those that didn’t—for the purpose of helping others generate even more possible experiments in this space.

Inference-time search

- To what extent can you learn complex inference-time search algorithms? Could you structure an architecture that would learn MCTS organically? How artificial does such an architecture have to be? How do factors like environment diversity, inductive biases, algorithmic range, statefulness, and hard-coded optimization impact this?
- Does modeling search algorithms produce search algorithms? If an RL agent is in an environment with another agent that is doing some hard-coded search and it has to model that agent, does it learn to implement search itself?
- What happens if you let a model not have constant computational complexity? Can it learn search better?
- Can you learn optimization/search algorithms even in non-RL settings? Can you get a purely predictive model to perform well zero-shot on an RL task?

Reward unidentifiability

- How do RL agents behave in environments with unidentifiable goals? What sorts of proxies tend to be favored over others? How do different inductive biases

affect this? What about adversarial training? Can it be empirically demonstrated that most learned objectives are pseudo-aligned?

- Can you definitively distinguish between capability generalization and objective generalization? What if you train a planner and then test it off-distribution?
- Can you distinguish between the different forms of proxy alignment? Can you get an instrumentally pseudo-aligned model? Can you replicate the cleaning robot examples from Risks from Learned Optimization in gridworlds?
- Can you produce approximate alignment if you constrain model capacity?
- What about suboptimality alignment? Can you create an environment with simple and complex strategies with different alignment properties and get a model to switch between them at test time?
- How does using a learned reward model vs. learned policy affect unidentifiability concerns? Can you distinguish between and disincentivize misaligned objectives if you have access to a learned reward model? What about if you just have a Q function?
- Can you build a meta-learner that optimizes for robustness? How does meta-learning help with building robust models? What about one that minimizes time complexity?
- Can you do adversarial training with pseudo-inputs or other relaxations?

Modeling of the training process

- What happens if you do RL where you give the agent access to the reward signal through a side channel in its observation? How does it affect robustness? What if it has to work to get the leaked info? What if it's noisy? How does using architectures with more planning affect this? How do inductive biases affect this?
- Can you get a system to model its own training process? Just the objective? Can you apply sufficient inductive biases with a complex enough objective that it has to use information in the environment to figure it out rather than learn it directly?
- Can you get a model to defect off-distribution in a way that involves no defection on-distribution?
- To what extent do models care about their performance across episodes? If there exists a side-channel which only increases next-episode performance, under what circumstances will a model exploit such a thing?

-
1. Though please do say that you got the idea from here and let me know about any results that you get. ↪

SSC Meetups Everywhere: Community Map on frontpage this month

To celebrate and help people find meetups for this year's [SSC meetups everywhere](#), we are pinning the big community map to the frontpage map again for the month of September (as we did last year).

If you want to get rid of the map, you can do so in your [account settings](#) by clicking the "hide frontpage map" checkbox.

We also just added functionality to allow you to add a pin for yourself to the map, and sign up for notifications for nearby events and groups (accessible via the [community page](#)). You now also have the ability to specify a threshold for other people who have added themselves to the map in a given radius, which will also cause you to get a notification as soon as that threshold gets passed (to help facilitate new meetups in areas that just reached critical mass).

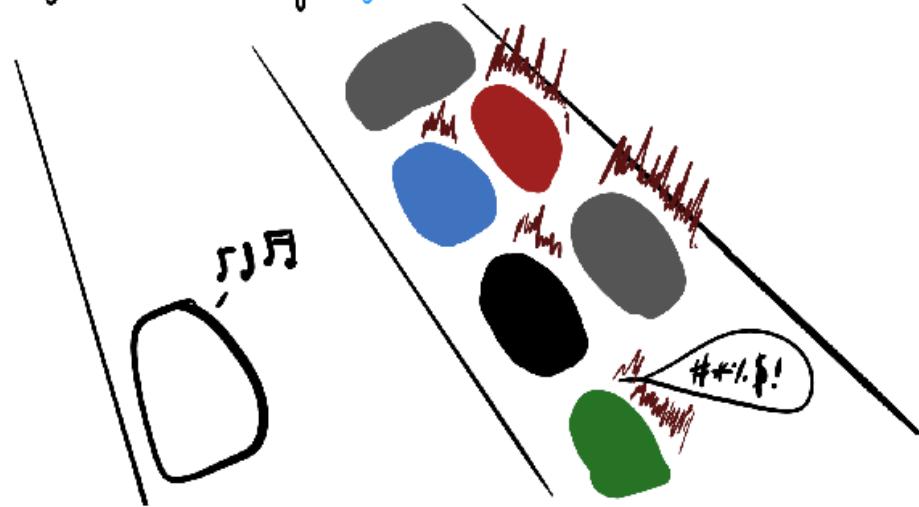
We will also be adding all the new groups and events that get posted to SSC to the meetup system.

Deducing Impact

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Impact is in the eye of the beholder.

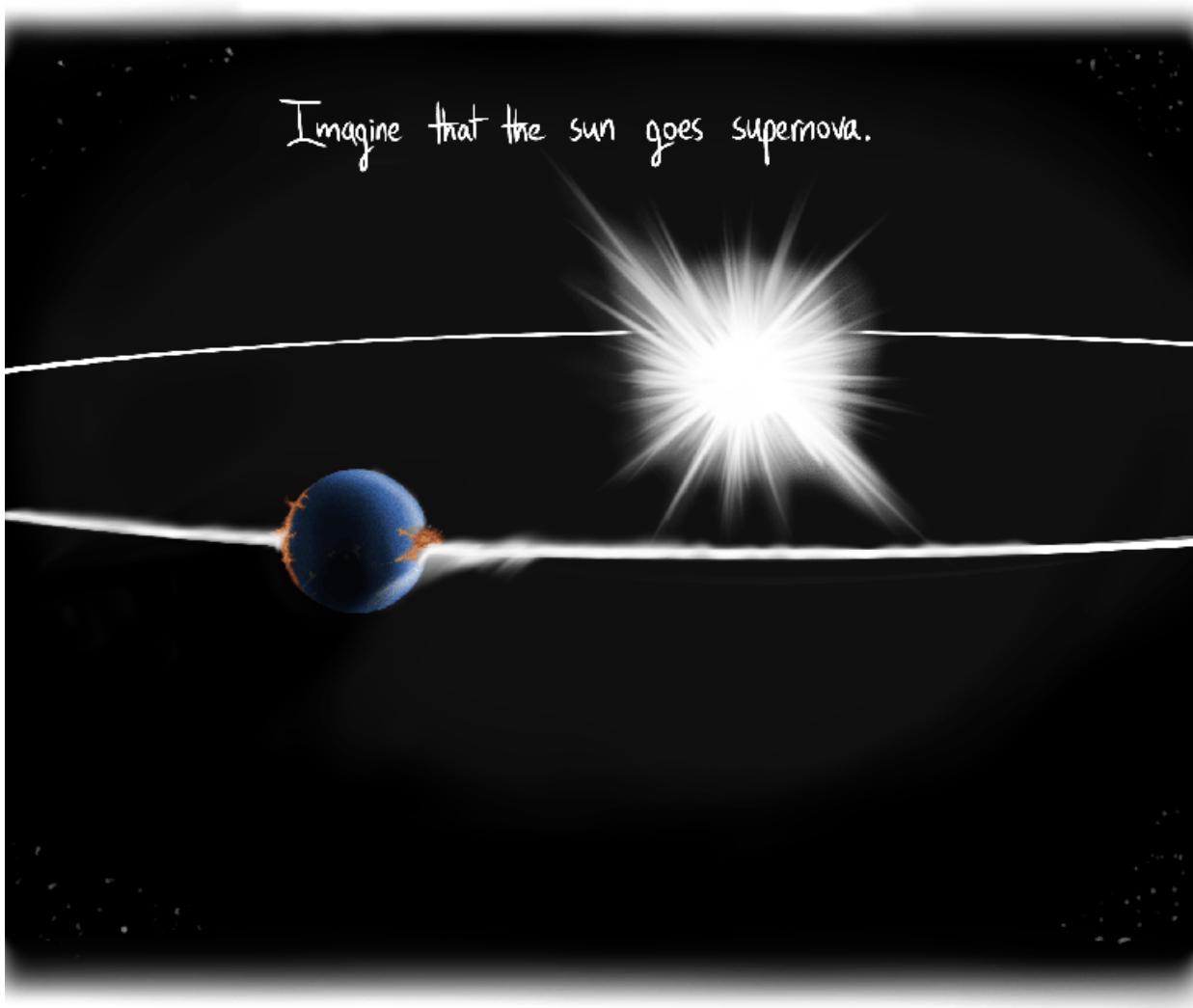
Traffic jams are vividly *big deals* to tardy commuters,



but everyone else doesn't really care.

This concept is important, so I'm going to present another zany situation.

Imagine that the sun goes supernova.



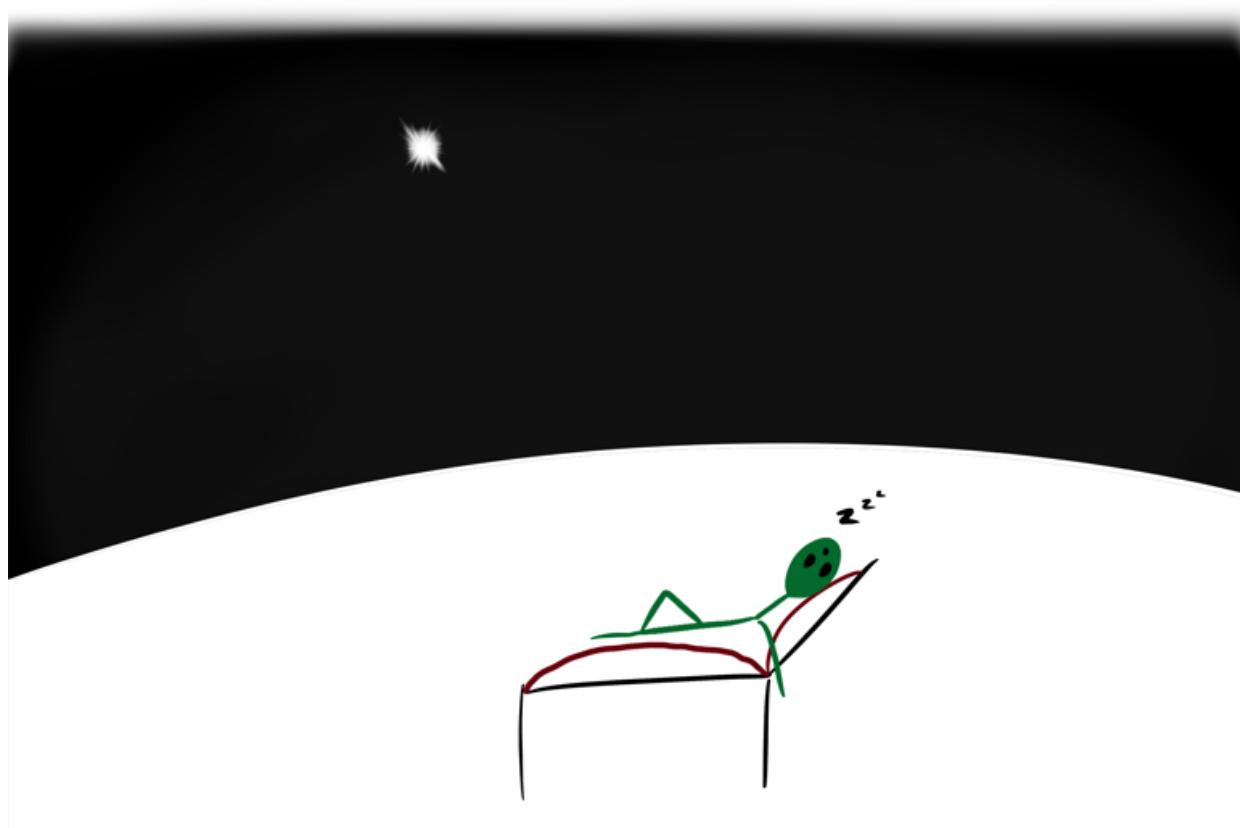
Now, being on -



Our sun is a main-sequence G-type star, it can't explode. Any energy input just decreases the volume of the hydrogen plasma, the Sun doesn't have a degenerate core that could be detonated. The Sun doesn't have enough mass to go supernova.

Yes, yes, thank you - everyone knows that.
It's just another weird situation I'm forcing on my readership.

Now, being on Earth here is objectively impactful because it matters to almost any agent in your shoes. However, whether this is a big deal to you depends on who and where you are.



Impact is Comparative

Suppose you grew up on an Earth where astrophysicists not only believe that stars less than eight solar masses can go supernova, but that the sun will — and soon.

Everyone knows it.

Everyone expects it to happen.

Everyone thinks it's unavoidable.

Including you.

Turns out, someone miscalculated; your calculations say it won't happen.

Saying this feels like a big deal is like calling a supernova "visible".

But to us, the sun not going supernova is... expected — zero impact.

You can't have it both ways — the sense of impact we experience has to be comparative — concerning some kind of difference, or change.



9

Our overarching goal is to deeply understand why capable AIs with **goals** are incentivized to catastrophically **impact** us, and then discuss an agent design which potentially avoids these incentives. The first step towards this end is, of course, understanding **impact**.

One rarely has the opportunity to try their hand at connecting the dots of an important insight. In AI alignment, I suspect that some important questions have yet to be properly posed, let alone solved. To get feedback and grow as a researcher, one must hone their reason on the known conquests of other fields.

This question of **impact** is not the important question, but it is an important question. It has both reasonable difficulty and an answer not yet widely circulated. Go ahead — try your hand.

6

Exercise: Deducing and informally describing why we think some things are **big deals**. Remember:

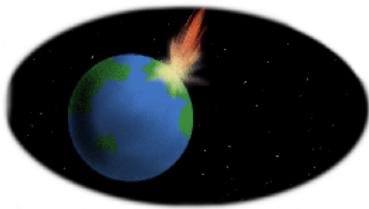
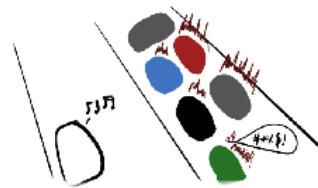
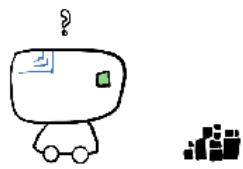
- Impact is relative both to what you value and to your vantage point.
- Part of impact is particular to agents like you, and part is objective.
- Impact is comparative.

Find the simple concept neatly explaining why things feel like **big deals** or not in the examples thus far.

The answer can be expressed in just one sentence of everyday language.

You have the benefit of knowing a solution exists.

You have fifteen minutes.



The solution comes in the next post! Feel free to discuss amongst yourselves.

Reminder: Your sentence should explain impact from all of the perspectives we discussed (from XYZ to humans).

Is Specificity a Mental Model?

This is Part IX of the [Specificity Sequence](#)

I've recently noticed a lot of smart people publishing lists of **mental models**. They're apparently having a moment:

- [Mental Models: The Best Way to Make Intelligent Decisions](#) by Shane Parrish
- [Mental Models I Find Repeatedly Useful](#) and [Super Thinking: The Big Book of Mental Models](#) by Gabriel Weinberg and Lauren McCann
- [Mental Models: Learn How to Think Better and Gain a Mental Edge](#) by James Clear
- [Mental Models](#) by Julian Shapiro
- [A Lesson on Elementary Worldly Wisdom](#), a 1994 talk by Charlie Munger about how he and Warren Buffet use a "latticework of models" to understand the world and make good investments. A couple decades early to the current party, but still.

There are hundreds of useful mental models to learn, such as "leverage", "social proof", "seizing the middle", and of course, "mental model". I want to help you be the very best, searching far and wide, teaching your brain to understand the power that's inside.



Gotta Catch 'Em All

We've been focusing nonstop on one (super)powerful mental model called the "[ladder of abstraction](#)", and seen it prove useful in a surprising variety of unrelated domains. The best mental models are the ones that have the largest number of applicability domains while also being the simplest and most compact. ☺

But despite all its usefulness, the ladder of abstraction doesn't appear in any list of mental models I've seen to date. The closest it's gotten is probably this entry from [Farnam Street's list](#) in the "Military & War" section:

Seeing the Front

One of the most valuable military tactics is the habit of "personally seeing the

front” before making decisions—not always relying on advisors, maps, and reports, all of which can be either faulty or biased.

Yes, advisors and maps and reports that tell you the reality on the ground may be “faulty or biased”, but there’s an even more fundamental problem: Their whole job is to slide the ground truth *up* the ladder of abstraction.

A report tells you that your enemy’s troops on the battlefield outnumber yours 2-to-1. Sounds like you should retreat, right? Not so fast. Let’s slide that down the ladder of abstraction by filling in more detail.



Rain clouds are gathering in the sky? The report might have neglected to mention that detail. When you realize it’s going to rain, you might think of a clever strategy that uses the rain to your advantage.

“Seeing the front” is an instance of “sliding down the ladder of abstraction”: you replace an abstract summarization of observations of the front with a lower-level data dump about the front (which happens to come from your own senses).

“Seeing the front” is absolutely a useful mental model to know; so are others like “[Value Prop Story](#)” and “[mind-anchor](#)”. But “[ladder of abstraction](#)” is even *more* useful because it lets you derive these and a bunch of other mental models for yourself.

Next post: [The Power to Draw Better](#)