Best of LessWrong: April 2019

1. Asymmetric Justice
2. The Forces of Blandness and the Disagreeable Majority
3. The Principle of Predicted Improvement
4. Degrees of Freedom
5. Literature Review: Distributed Teams
6. How do people become ambitious?
7. 1960: The Year The Singularity Was Cancelled
8. Helen Toner on China, CSET, and AI
9. Best reasons for pessimism about impact of impact measures?
10. Book review: The Sleepwalkers by Arthur Koestler
11. Any rebuttals of Christiano and AI Impacts on takeoff speeds?
12. Where to Draw the Boundaries?
13. Robin Hanson on Simple, Evidence Backed Models
14. Why does category theory exist?
15. Counterspells
16. The Hard Work of Translation (Buddhism)
17. [Answer] Why wasn't science invented in China?
18. What are some good examples of incorrigibility?
19. LW Update 2019-04-02 – Frontpage Rework
20. "Everything is Correlated": An Anthology of the Psychology Debate
21. Pecking Order and Flight Leadership
22. Alignment Newsletter One Year Retrospective
23. Prompts for eliciting blind spots/bucket errors/bugs
24. Conditional revealed preference
25. Liar Paradox Revisited
26. Against Street Epistemology
27. Machine Pastoralism
28. The Stack Overflow of Factored Cognition
29. Experimental Open Thread April 2019: Socratic method
30. Recent updates to gwern.net (2017–2019)
31. Could waste heat become an environment problem in the future (centuries)?
32. Crypto quant trading: Intro
33. Evidence other than evolution for optimization daemons?
34. The Simple Solow Model of Software Engineering
35. When is rationality useful?
36. Moving to a World Beyond “$p < 0.05$”
37. Rationality made me less bad at Mario Kart 8
38. [April Fools] User GPT2 is Banned
39. Strategic implications of AIs' ability to coordinate at low cost, for example by merging
40. Why is multi worlds not a good explanation for abiogenesis
41. Moral Weight Doesn't Accrue Linearly
42. Excerpts from a larger discussion about simulacra
43. Value Learning is only Asymptotically Safe
44. IRL 7/8: Generalizing human-robot cooperation: Cooperative IRL
45. IRL 5/8: Maximum Causal Entropy IRL
46. [AN #54] Boxing a finite-horizon AI system to keep it unambitious
47. Many maps, Lightly held
48. Scrying for outcomes where the problem of deepfakes has been solved
49. Alignment Newsletter #52
50. Quantitative Philosophy: Why Simulate Ideas Numerically?

Best of LessWrong: April 2019

1. Asymmetric Justice
2. The Forces of Blandness and the Disagreeable Majority
3. The Principle of Predicted Improvement
4. Degrees of Freedom
5. Literature Review: Distributed Teams
6. How do people become ambitious?
7. 1960: The Year The Singularity Was Cancelled
8. Helen Toner on China, CSET, and AI
9. Best reasons for pessimism about impact of impact measures?
10. Book review: The Sleepwalkers by Arthur Koestler
11. Any rebuttals of Christiano and AI Impacts on takeoff speeds?
12. Where to Draw the Boundaries?
13. Robin Hanson on Simple, Evidence Backed Models
14. Why does category theory exist?
15. Counterspells
16. The Hard Work of Translation (Buddhism)
17. [Answer] Why wasn't science invented in China?
18. What are some good examples of incorrigibility?
19. LW Update 2019-04-02 – Frontpage Rework
20. "Everything is Correlated": An Anthology of the Psychology Debate
21. Pecking Order and Flight Leadership
22. Alignment Newsletter One Year Retrospective
23. Prompts for eliciting blind spots/bucket errors/bugs
24. Conditional revealed preference
25. Liar Paradox Revisited
26. Against Street Epistemology
27. Machine Pastoralism
28. The Stack Overflow of Factored Cognition
29. Experimental Open Thread April 2019: Socratic method
30. Recent updates to gwern.net (2017–2019)
31. Could waste heat become an environment problem in the future (centuries)?
32. Crypto quant trading: Intro
33. Evidence other than evolution for optimization daemons?
34. The Simple Solow Model of Software Engineering
35. When is rationality useful?
36. Moving to a World Beyond “$p < 0.05$”
37. Rationality made me less bad at Mario Kart 8
38. [April Fools] User GPT2 is Banned
39. Strategic implications of AIs' ability to coordinate at low cost, for example by merging
40. Why is multi worlds not a good explanation for abiogenesis
41. Moral Weight Doesn't Accrue Linearly
42. Excerpts from a larger discussion about simulacra
43. Value Learning is only Asymptotically Safe
44. IRL 7/8: Generalizing human-robot cooperation: Cooperative IRL
45. IRL 5/8: Maximum Causal Entropy IRL
46. [AN #54] Boxing a finite-horizon AI system to keep it unambitious
47. Many maps, Lightly held
48. Scrying for outcomes where the problem of deepfakes has been solved

49. Alignment Newsletter #52
50. Quantitative Philosophy: Why Simulate Ideas Numerically?

Asymmetric Justice

Related and required reading in life (ANOIEAEIB): The Copenhagen Interpretation of Ethics

Epistemic Status: Trying to be minimally judgmental

Spoiler Alert: Contains minor mostly harmless spoiler for The Good Place, which is the best show currently on television.

The Copenhagen Interpretation of Ethics (in parallel with the similarly named one in physics) is as follows:

> The Copenhagen Interpretation of Ethics says that when you observe or interact with a problem in any way, you can be blamed for it. At the very least, you are to blame for not doing *more*. Even if you don't make the problem worse, even if you make it slightly better, the ethical burden of the problem falls on you as soon as you observe it. In particular, if you interact with a problem and benefit from it, you are a complete monster. I don't subscribe to this school of thought, but it seems pretty popular.

I don't say this often, but seriously, read the whole thing.

I do not subscribe to this interpretation.

I believe that the majority of people *effectively* endorse this interpretation. I do not think they endorse it consciously or explicitly. But they act as if it is true.

Another aspect of this same phenomenon is how most people view justice.

Almost everyone agrees justice is a sacred value. That it is good and super important. Justice is one of the few universally agreed upon goals of government. Justice is one of the eight virtues of the avatar. Justice is up there with truth and the American way. No justice, no peace.

But what is justice? Or rather, to avoid going too deeply into an infinitely complex philosophical debate millenniums or eons old, how do most people instinctively model justice in broad terms?

In a conversation last night, this was offered to me (I am probably paraphrasing due to bad memory, but it's functionally what was said), and seems common: Justice is giving appropriate punishment to those who have taken bad action.

I asked whether, in this person's model, the actions needed to be *bad* in order to be relevant to justice. This prompted pondering, after which the reply was that yes, that was how their model worked.

I then asked whether rewarding a good action counted as justice, or failing to do so counted as injustice, using the example of saving someone's life going unrewarded.

We can consider three point-based justice systems.

In the asymmetric system, when bad action is taken, bad action points are accumulated. Justice punishes in proportion to those points to the extent possible.

Each action is assigned a non-negative point total.

In the symmetric system, when *any* action is taken, *good or bad,* points are accumulated. This can be and often is zero, is negative for bad action, positive for good action. Justice consists of punishing negative point totals and rewarding positive point totals.

In what we will call the Good Place system (Spoiler Alert for Season 1), when any action is taken, good or bad, points are accumulated as in the symmetric system. But there's a catch (which is where the spoiler comes in). If you take actions with good consequences, you only get those points *if your motive was to do good.* When a character attempts to score points by holding open doors for people, they fail to score any points because they are gaming the system. Gaming the system isn't allowed.

Thus, if one takes action even under the best of motives, one fails to capture much of the gains from such action. Second or higher order benefits, or surprising benefits, that are real but unintended, will mostly not get captured.

The opposite is not true of actions with bad consequences. You lose points for bad actions whether or not you *intended* to be bad. It is *your* responsibility to check yourself before you wreck yourself.

When (Spoiler Alert for Season 3) an ordinary citizen buys a tomato from a supermarket, they are revealed to have *lost twelve points* because the owner of the tomato company was a bad guy and the company used unethical labor practices. Life has become too complicated to be a good person. Thus, since the thresholds never got updated, no one has made it into The Good Place for centuries.

The asymmetric system is *against action.* Action is bad. Inaction is good. Surprisingly large numbers of people actually believe this. It is good *to be you,* but bad *to do anything.*

The asymmetric system is not against *every* action. This is true. But effectively, it is. Some actions are bad, some are neutral. Take enough actions, even with the best of intentions, even with fully correct knowledge of what is and is not bad, and mistakes will happen.

So any individual, any group, any company, any system, any anything, that takes action, is therefore bad.

The law by design works that way, too. There are increasingly long and complex lists of actions which are illegal. If you break the law, and anyone who does things will do so by accident at some point, you can be prosecuted. You are then prosecuted for *the worst thing they can pin on you*. No amount of other good deeds can do more than mitigate. Thus, any sufficiently rich investigation will judge any of us who regularly take meaningful action to be bad.

If you can be sued for the bad consequences of a medical procedure, potentially for ruinous amounts, but cannot collect most of the huge benefits of successful procedures, you will engage in defensive medicine. Thus, lots of defensive medicine. Because justice.

If, as was done in the past, the engineer and his family are forced to sleep under the bridge after it is built, so that they will be killed if it falls down, you can be damn sure

they’re going to build a safe bridge. But you’d better want to pay for a fully bulletproof bridge before you do that.

Skin in the game is necessary. That means both being at risk, and collecting reward. Too often we assign risk without reward.

If one has a system whereby people are judged only by their bad actions, or by their *worst single* action, what you have is *a system that condemns and is against all action*.

Never tweet.

Also see privacy and blackmail.

The symmetric system is in favor of action. If no one ever took any action, we would not have nice things and also all die. If people generally took fewer actions, we would have less nice things and be worse off. If one gets full credit for the good and bad consequences of one’s actions, we will provide correct incentives to encourage action.

This, to me, is also justice.

A symmetric system can still count bad consequences as *larger* than similar good consequences to a large extent (e.g. saving nine people from drowning does not give one enough credits to murder a tenth), and we can punish locally bad intent on top of direct consequences, without disturbing this. Action is on net a very good thing.

The Good Place system works well for *simple* actions with mostly *direct* consequences. One then, under normal circumstances, gets credit for the good and the bad. It also has a great feature, which is that it *forces the action via a high required threshold.* You need a *lot* of points to pass a binary evaluation when you die. Sitting around doing nothing is a very bad idea.

The problem comes in when there are *complex indirect* consequences that are hard to fully know or observe.

Some of the indirect consequences of buying a tomato are good. You don’t get credit for those unless you knew about them, because all you were trying to do was buy a tomato. Knowing about them is possible in theory, but expensive, and doesn’t make them better. It only makes you know about them, which only matters to the extent that it changes your decisions.

Some of the indirect consequences of buying a tomato are bad. You lose those points.

Thus, when you buy a tomato and thus another customer can’t buy a tomato, you get docked. But when you buying a tomato increases the store’s estimated demand for tomatoes, so they order more and don’t run out next week, and a customer gets to buy one (and the store stays in business to provide even more tomatoes), you don’t get rewarded.

Better to not take the shopping action.

No wonder people make seemingly absurdist statements like “there is no ethical consumption under capitalism.”

Under this philosophy, there is no ethical action under complexity. Period.

I get that complexity is bad. But this is ridiculous.

Compare to the Copenhagen Interpretation of Ethics. If one interacts with a compact, isolated problem, such as a child drowning in a pond, one can reasonably do all one could do, satisfying one's requirements. If one interacts with or observes a non-compact, non-isolated problem, such as third world poverty, you are probably Mega-Hitler. You cannot both be a good person and have slack.

As a young child, I read the book Be a Perfect Person in Just Three Days. Spoiler alert, I guess? The protagonist is given a book with instructions on how to be a perfect person. The way to do so is to take progressively less action. First day you take symbolic action, wearing broccoli around your neck. Second day you take inaction, by fasting. Third day, *you do nothing at all except drink weak tea and go to the bathroom.*

That makes you 'perfect.'

Because perfect means a score of exactly zero points.

Asymmetric systems of judgment are systems for opposing all action.

The Forces of Blandness and the Disagreeable Majority

There are a few data points that have been making me see “the discourse” differently lately.

1. Large Majorities Dislike Political Correctness.

That’s the title of this Atlantic article that came out in October, and is based on this study from the think tank More in Common which opposes political polarization.

The results of the 8000-person poll of a nationally-representative sample of Americans are pretty striking. About 80% of Americans think “political correctness is a problem”; and even when you restrict to self-identified liberals, Democrats, or people of color, large majorities agree with the statement. The study identifies “progressive activists” (8% of Americans) as a younger, more extreme, more educated, more politically active left-wing cluster, and even within this cluster, a full 25% agree with “political correctness is a problem.”

And lots of people who agree with statements about hate speech being bad, white people starting out with advantages in life, sexual harassment being a problem, etc, *also* think political correctness is a problem.

Being “politically incorrect” isn’t just a white thing, a male thing, or even a conservative thing. It’s a *hugely common thing*.

2. Support for free speech is common, and *growing*, not shrinking. And it’s *not* the most left-wing people who most oppose free speech, but the *moderate* liberals.

Political scientist Justin Murphy has done studies about this, based on the General Social Survey, a large poll on social attitudes that’s been running for decades.

Since the 1970’s, Americans have become *more* tolerant of allowing people with controversial views to speak in public — communists, people proposing military coups, homosexuals, and opponents of “all churches and religions.” Racism is the exception to the rule — people *haven’t* become more tolerant of racist speech, even as they *have* become more tolerant of other varieties of speech.

Keep in mind that legal censorship and centralization of political speech were *way* more prevalent in mid-20th century America than they are today. Cable television networks didn’t exist till the 1970’s. The Fairness Doctrine didn’t end until 1987. Satellite radio, which allowed obscene language that was regulated on conventional radio and television, only began in 1988, Fox News was founded in 1996, and, of course, the blogosphere didn’t really begin until the early 2000’s.

Murphy notes that “extreme liberals” are consistently the *most* supportive of permitting controversial speech, and that in fact they have *increased* their rates of tolerating even racist speech. People who rate themselves as “moderately liberal” and “slightly liberal”, however, have sharply declined in their willingness to tolerate racist speech. If there’s been a “backlash against free speech”, it’s on the *moderate* left, not the far left.

3. Calls for speech restrictions often come from moderates.

Things like this essay by Renee diResta, which I found chilling — a call for social media to be actively regulated *by the US military,* which says we should treat people spreading opinions that weaken trust in "the legitimacy of government, the persistence of societal cohesion, even our ability to respond to the impending climate crisis" as "digital combatants." DiResta says, "More authoritarian regimes, by contrast, would simply turn off the internet. An admirable commitment to the principle of free speech in peace time turns into a sucker position against adversarial psy-ops in wartime."

Who is DiResta? She's a writer, technologist, adviser to Congress and the State Department, and the director of research at something called New Knowledge, a firm offering corporations a new kind of service: using algorithms to bury social media scandals that would make them look bad.

In other words, she's an influential moderate; well-connected in corporate and government worlds, and *very* troubled by the crisis of declining trust in traditional institutions that the open Internet has enabled.

An Alternative Paradigm: Moderate, Measured Elites vs. The Chaotic, Offensive Populace

What if "free speech" vs. "restricted speech" isn't a right-vs.-left thing at all?

Lots of people, who are by no means all political conservatives, want the right to say offensive things.

Verbal conflict just isn't that big a problem to most people, apparently. And how likely you are to violate vs. observe verbal taboos varies a lot based on personality and socioeconomic class.

Swearing is an interesting example of a verbal taboo that's not especially politicized. Socially low-ranking people swear more. Swearing is negatively correlated with agreeableness. Men swear more than women. Swearing is commonly associated with being working-class, though I haven't found published evidence of this. Swearing is "inappropriate" in office settings, religious settings, or whenever we're expected to be formal or respectful.

It's often corporate caution that drives speech codes that restrict political controversy, obscenity, and muckraking/whistleblowing. It's not just racist or far-right opinions that get silenced; media and social-media corporations worry about offending prudes, homophobes, Muslim extremists, the Chinese government, the US military, etc, etc.

Some people clearly *do* have strong ideological opinions about what speech they want to see allowed vs. banned, but I don't see that as the main driver of what rules actually get put into place. What I think is going on is that *decisionmakers in media and PR, and corporate and government elites generally, have a lower tolerance for verbal conflict and taboo violations than the typical individual.*

The growth of lots and lots of outlets for more "unofficial" or "raw" self-expression — blogs, yes, but before that cable TV and satellite radio, and *long* before that, the culture of "journalism" in 18th century America where every guy with a printing press could publish a "newspaper" full of opinions and scurrilous insults — tends to go along with more rudeness, more cursing, more sexual explicitness, more political

extremism in *all* directions, more “trashy” or “lowest common denominator” media, more misinformation and “dumbing down”, but also some innovative/intellectual “niche” media.

Chaos is a centrifugal force; it increases the chance of *any* unexpected outcome. Good things, bad things, existential threats, brilliant ideas, and a *lot* of weird, gross, and disturbing stuff.

Some people like parts of that (it’s hard to like *everything* about chaos), and others find even a little chaos threatening. The most passionate opponents of chaos are likely to be powerful, since change can only knock them off their pedestals.

I think we’re currently in an era of *unusually large amounts of free speech that elites are starting to get spooked by and defend against.* Most people have high, perhaps even growing, tolerance for controversy and offense, but some find it unacceptable, and these people are disproportionately influential.

The Principle of Predicted Improvement

I made a conjecture I think is cool. Mark Sellke proved it. I don't know what else to do with it, so I will explain why I think it's cool and give the proof here. Hopefully, you will think it's cool, too.

Suppose we are trying to assign as much probability as possible to whichever of several hypotheses is true. The law of conservation of expected evidence tells us that for any hypothesis, we should expect to assign the same probability to that hypothesis after observing a test result that we assign to it now. Suppose that H takes values h_i.

We can express the law of conservation of expected evidence as, for any fixed h_i:

$$E[P(H = h_i \mid D)] = P(H = h_i)$$

In English this says that the probability we should expect to assign to h_i after observing the value of D equals the probability we assign to h_i before we observe the value of D.

This law raises a question. If all I want is to assign as much probability to the true hypothesis as possible, and I should expect to assign the same probability I currently assign to each hypothesis after getting a new piece of data, why would I ever collect more data? A. J. Ayer pointed out this puzzle in *The Conception of Probability as a Logical Relation* (I unfortunately cannot find a link). I. J. Good solved Ayer's puzzle in *On the Principle of Total Evidence*. Good shows that if I need to act on a hypothesis, the expected value of gaining an extra piece of data is always greater than or equal to the expected value of not gaining that new piece of data. Although there is nothing wrong with Good's solution, I found it somewhat unsatisfying. Ayer's puzzle is purely epistemic, and while there is nothing wrong with a pragmatic solution to an epistemic puzzle, I still felt that there should be a solution that makes no reference to acts or utility at all.

Herein I present a theorem that I think constitutes such a solution. I have decided to call it the principle of predicted improvement (PPI):

$$E[P(H \mid D)] \geq E[P(H)]$$

In English the theorem says that the probability we should expect to assign to the true value of H after observing the true value of D is greater than or equal to the expected probability we assign to the true value of H before observing the value of D. This inequality is strict when H and D are not independent. In other words, you should

predict that your epistemic state will improve (e.g. you will assign more probability to the truth) after making any relevant observation.

This is a solution to Ayer's puzzle because it says that I should always expect to assign more probability to the true hypothesis after making a relevant observation. It is a purely epistemic solution because it makes no reference to acts or utility. So long as I want to assign more probability to the true hypothesis than I currently do, I should want to make relevant observations.

Importantly, this is completely consistent with the law of conservation of expected evidence. Although for any particular hypothesis I should expect to assign it the same probability after performing a test that I do now, I should also expect to assign more probability to whichever hypothesis is actually true.

Aside from being a solution to Ayer's puzzle, the PPI is cool just because it tells you that you should expect to assign more probability to the truth as you observe stuff.

There is a similar more well known theorem from information theory that my friend Alex Davis showed me:

$$E\left[E\left[-\log\left(P(H \mid D)\right) \mid D\right]\right] \leq E\left[-\log\left(P(H)\right)\right]$$

In English this says that you should expect the entropy of your distribution to go down after you make an observation. If we use the law of iterated expectation, multiply both sides by minus one, and reverse the inequality, we get something that looks a lot like the PPI:

$$E\left[\log\left(P(H \mid D)\right)\right] \geq E\left[\log\left(P(H)\right)\right]$$

It does not imply PPI in any obvious way because we can have two distributions such that one is higher in expected negentropy but lower in expected probability assigned to the truth, and vice versa. They are similar theorems in that one says you should predict that the probability assigned to the true outcome will be higher after an observation, while the other says you should predict that the log probability will be higher. They are different in that they use different measures of confidence.

The advantage of the PPI is that it is phrased in the same terms as Ayer's puzzle: probabilities rather than log probabilities. I also claim that the PPI is easier to read and interpret, so it might be pedagogically useful to teach it before teaching that expected entropy after an observation is less than or equal to current entropy.

Anyway, here's Sellke's proof.

Proof:

We want to show

$$E[P(H \mid D)] \geq E[P(H)].$$

Let's say that H takes values (h_i) and D takes values (d_j). The left hand side is

$$\sum_{i,j} P[h_i \mid d_j] P[h_i \wedge d_i].$$

The right hand side is

$$\sum_i P[h_i]^2.$$

The left hand side is equivalent to

$$\sum_{i,j} \frac{P[h_i \wedge d_j]^2}{P[d_j]} = \sum_i \sum_j \frac{P[h_i \wedge d_j]^2}{P[d_j]}.$$

Titu's lemma: for any sequences of a's and b's

$$\sum_j \left(\frac{a_j}{b_j}\right)^2 \geq \frac{\left(\sum_j a_j\right)^2}{\sum_j b_j}.$$

If we apply this to each $\sum_j$ then we just get $P[h_i]^2$. This is because:

$$\sum_j P[h_i \wedge d_j] = P[h_i]$$

and

$$\sum_j P[d_j] = 1.$$

For each fixed i, set:

$$a_j = a_{i,j} = P[h_i \wedge d_j]$$

and

$$b_j = P[d_j].$$

To conclude, for each fixed i we have

$$\sum_j \frac{P[h_i \wedge d_j]^2}{P[d_j]} \geq \frac{\left(\sum_j P[h_i \wedge d_j]\right)^2}{\sum_j P[d_j]} = P[h_i]^2$$

and hence

$$\sum_i \sum_j \frac{P[h_i \wedge d_j]^2}{P[d_j]} \geq \sum_i P[h_i]^2.$$

Equality:

Here we explain why the only equality case is when H and E are independent.

Titu's Lemma is an equality iff the two vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ are parallel, that is, if there exists a constant λ such that $a_i = \lambda b_i$ for all i. If we translate this equality condition over to our application of Titu's Lemma above, we see that our proof preserves equality if and only if there exist constants $\lambda_1, \ldots, \lambda_n$ such that $P[h_i \wedge d_j] = \lambda_i \cdot P[d_j]$. (We applied Titu once for each value of i, so we need a λ_i value for each inequality to be an equality. But these λ_i's can be different.)

Now if we sum over j there we get

$$\sum_j P[h_i \wedge d_j] = \lambda_i \cdot \sum_j P[d_j]$$

and so

$$P[h_i] = \lambda_i.$$

Plugging this back in, we see that equality is true iff

$$P[h_i \wedge d_j] = P[h_i] \cdot P[d_j]$$

which is equivalent to independence of H and E, or 0 mutual information.

Degrees of Freedom

Something I've been thinking about for a while is the dual relationship between optimization and indifference, and the relationship between both of them and the idea of freedom.

Optimization: "Of all the possible actions available to me, which one is best? (by some criterion). Ok, I'll choose the best."

Indifference: "Multiple possible options are equally good, or incommensurate (by the criterion I'm using). My decision algorithm equally allows me to take any of them."

Total indifference between all options makes optimization impossible or vacuous. An optimization criterion which assigns a total ordering between all possibilities makes indifference vanishingly rare. So these notions are dual in a sense. Every dimension along which you optimize is in the domain of optimization; every dimension you leave "free" is in the domain of indifference.

Being "free" in one sense can mean "free to optimize". I choose the outcome that is best according to an *internal* criterion, which is not blocked by *external* barriers. A limit on freedom is a constraint that keeps me away from my favorite choice. Either a natural limit ("I would like to do that but the technology doesn't exist yet") or a man-made limit ("I would like to do that but it's illegal.")

There's an ambiguity here, of course, when it comes to whether you count "I would like to do that, but it would have a consequence I don't like" as a limit on freedom. Is that a barrier blocking you from the optimal choice, or is it simply another way of saying that it's not an optimal choice after all?

And, in the latter case, isn't that basically equivalent to saying there *is* no such thing as a barrier to free choice? After all, "I would like to do that, but it's illegal" is effectively the same thing as "I would like to do that, but it has a consequence I don't like, such as going to jail." You can get around this ambiguity in a political context by distinguishing natural from social barriers, but that's not a particularly principled distinction.

Another issue with freedom-as-optimization is that it's compatible with quite tightly constrained behavior, in a way that's not consistent with our primitive intuitions about freedom. If you're only "free" to do the optimal thing, that can mean you are free to do only *one* thing, all the time, as rigidly as a machine. If, for instance, you are only free to "act in your own best interests", you don't have the option to act *against* your best interests. People in real life *can* feel constrained by following a rigid algorithm even when they agree it's "best"; "but what if I want to do something that's *not* best?" Or, they can acknowledge they're free to do what they choose, but are dismayed to learn that their choices are "dictated" as rigidly by habit and conditioning as they might have been by some human dictator.

An alternative notion of freedom might be freedom-as-arbitrariness. Freedom in the sense of "degrees of freedom" or "free group", derived from the intuition that freedom means *breadth of possibility* rather than *optimization power*. You are only free if you *could* equally do any of a number of things, which ultimately means something like indifference.

This is the intuition behind claims like Viktor Frankl's: "Between stimulus and response there is a space. In that space is our power to choose a response. In our response lies our growth and our freedom." If you always respond automatically to a given stimulus, you have only one choice, and that makes you unfree in the sense of "degrees of freedom."

Venkat Rao's concept of freedom is pretty much this freedom-as-arbitrariness, with some more specific wrinkles. He mentions degrees of freedom ("dimensionality") as well as "inscrutability", the inability to predict one's motion from the outside.

Buddhists also often speak of freedom more literally in terms of indifference, and there's a very straightforward logic to this; you can only choose equally between A and B if you have been "liberated" from the attractions and aversions that *constrain* you to choose A over B. Those who insist that Buddhism is compatible with a fairly normal life say that after Buddhist practice you still *will* choose systematically most of the time — your utility function cannot fully flatten if you act like a living organism — but that, like Viktor Frankl's ideal human, you will be able to reflect with equinamity and *consider* choosing B over A; you will be more "mentally flexible." Of course, some Buddhist texts simply say that you become *actually* indifferent, and that sufficient vipassana meditation will make you indistinguishable from a corpse.

Freedom-as-indifference, I think, is lurking behind our intuitions about things like "rights" or "ownership." When we say you have a "right" to free speech — even a right bounded with certain limits, as it of course always is in practice — we mean that within those limits, you may speak *however you want*. Your rights define a space, within which you may behave arbitrarily. Not optimally. A right, if it's not to be vacuous, *must* mean the right to behave "badly" in some way or other. To own a piece of property means that, within whatever limits the concept of ownership sets, you may make use of it in any way you like, even in suboptimal ways.

This is very clearly illustrated by Glen Weyl's notion of radical markets, which neatly disassociates two concepts usually both considered representative of free-market systems: ownership and economic efficiency. To own something just *is* to be able to hang onto it even when it is economically inefficient to do so. As Weyl says, "property is monopoly." The owner of a piece of land can sit on it, making no improvements, while holding out for a high price; the owner of intellectual property can sit on it without using it; in exactly the same way that a monopolist can sit on a factory and depress output while charging higher prices than he could get away with in a competitive market.

For better or for worse, rights and ownership define spaces in which you can destroy value. If your car was subject to a perpetual auction and ownership tax as Weyl proposes, bashing your car to bits with a hammer would cost you even if you didn't personally need a car, because it would hurt the rental or resale value and you'd still be paying tax. On some psychological level, I think this means you couldn't feel fully *secure* in your possessions, only probabilistically likely to be able to provide for your needs. You only truly own what you have a right to wreck.

Freedom-as-a-space-of-arbitrary-action is also, I think, an intuition behind the fact that society (all societies, but the US more than other rich countries, I think) is shaped by people's desire for more discretion in decisionmaking as opposed to transparent rubrics. College admissions, job applications, organizational codes of conduct, laws and tax codes, all are designed deliberately to allow ample discretion on the part of

decisionmakers rather than restricting them to following “optimal” or “rational”, simple and legible, rules. Some discretion is necessary to ensure good outcomes; a wise human decisionmaker can always make the right decision in some hard cases where a mechanical checklist fails, simply because the human has more cognitive processing power than the checklist. This phenomenon is as old as Plato’s *Laws* and as current as the debate over algorithms and automation in medicine. However, what we observe in the world is *more* discretion than would be necessary, for the aforementioned reasons of cognitive complexity, to generate socially beneficial outcomes. We have discretion that enables corruption and special privileges in cases that pretty much nobody would claim to be ideal — rich parents buying their not-so-competent children Ivy League admissions, favored corporations voting themselves government subsidies. Decisionmakers want the “freedom” to make illegible choices, choices which would look “suboptimal” by naively sensible metrics like “performance” or “efficiency”, choices they would prefer not to reveal or explain to the public. Decisionmakers feel trapped when there’s too much “accountability” or “transparency”, and prefer a wider sphere of discretion. Or, to put it more unfavorably, they want to be free to destroy value.

And this is true at an individual psychological level too, of course — we want to be free to “waste time” and resist pressure to account for literally everything we do. Proponents of optimization insist that this is simply a failure mode from picking the wrong optimization target — rest, socializing, and entertainment are *also* needs, the optimal amount of time to devote to them isn’t zero, and you don’t have to consider personal time to be “stolen” or “wasted” or “bad”, you can, in principle, legibilize your entire life including your pleasures. Anything you wish you could do “in the dark”, off the record, you could also do “in the light,” explicitly and fully accounted for. If your boss uses “optimization” to mean overworking you, the problem is with your boss, not with optimization per se.

The freedom-as-arbitrariness impulse in us is skeptical.

I see optimization and arbitrariness everywhere now; I see intelligent people who more or less take one or another as ideologies, and see them as *obviously* correct.

Venkat Rao and Eric Weinstein are partisans of arbitrariness; they speak out in favor of “mediocrity” and against “excellence” respectively. The rationale being, that being highly optimized at some widely appreciated metric — being very intelligent, or very efficient, or something like that — is often less valuable than being *creative*, generating something in a part of the world that is “dark” to the rest of us, that is not even on our map as something to value and thus appears as lack of value. Ordinary people being “mediocre”, or talented people being “undisciplined” or “disreputable”, may be more creative than highly-optimized “top performers”.

Robin Hanson, by contrast, is a partisan of optimization; he speaks out against bias and unprincipled favoritism and in favor of systems like prediction markets which would force the “best ideas to win” in a fair competition. Proponents of ideas like radical markets, universal basic income, open borders, income-sharing agreements, or smart contracts (I’d here include, for instance, Vitalik Buterin) are also optimization partisans. These are legibilizing policies that, if optimally implemented, can always be Pareto improvements over the status quo; “whatever degree of wealth redistribution you prefer”, proponents claim, “surely it is better to achieve it in whatever way results in the least deadweight loss.” This is the very reason that they are *not* the policies that public choice theory would predict would emerge naturally in governments. Legibilizing policies allow little scope for discretion, so they don’t let policymakers give

illegible rewards to allies and punishments to enemies. They reduce the scope of the "political", i.e. that which is negotiated at the personal or group level, and replace it with an impersonal set of rules within which individuals are "free to choose" but not very "free to behave arbitrarily" since their actions are transparent and they must bear the costs of being in full view.

Optimization partisans are against weakly enforced rules — they say "if a rule is good, enforce it consistently; if a rule is bad, remove it; but selective enforcement is just another word for favoritism and corruption." Illegibility partisans say that weakly enforced rules are the only way to incorporate valuable information — precisely that information which enforcers *do not feel they can make explicit*, either because it's controversial or because it's too complex to verbalize. "If you make everything explicit, you'll dumb everything in the world down to what the stupidest and most truculent members of the public will accept. Say goodbye to any creative or challenging innovations!"

I see the value of arguments on both sides. However, I have positive (as opposed to normative) opinions that I don't think everybody shares. I think that the world I see around me is moving in the direction of *greater arbitrariness* and has been since WWII or so (when much of US society, including scientific and technological research, was organized along military lines). I see arbitrariness as a thing that arises in "mature" or "late" organizations. Bigger, older companies are more "political" and more monopolistic. Bigger, older states and empires are more "corrupt" or "decadent."

Arbitrariness has a tendency to protect those in power rather than out of power, though the correlation isn't perfect. Zones that protect your ability to do "whatever" you want without incurring costs (which include zones of privacy or property) are *protective, conservative* forces — they allow people *security*. This often means protection for those who already have a lot; arbitrariness is often "elitist"; but it can also protect "underdogs" on the grounds of tradition, or protect them by shrouding them in secrecy. (Scott thought "illegibility" was a valuable defense of marginalized peoples like the Roma. Illegibility is not *always* the province of the powerful and privileged.) No; the people such zones of arbitrary, illegible freedom systematically *harm* are those who *benefit from increased accountability and revealing of information*. Whistleblowers and accusers; those who expect their merit/performance is good enough that displaying it will work to their advantage; those who call for change and want to display information to justify it; those who are newcomers or young and want a chance to demonstrate their value.

If your intuition is "you don't know me, but you'll like me if you give me a chance" or "you don't know *him*, but you'll be horrified when you find out what he did", or "if you gave me a chance to explain, you'd agree", or "if you just let me compete, I bet I could win", then you want more optimization.

If your intuition is "I can't explain, you wouldn't understand" or "if you knew what I was *really* like, you'd see what an impostor I am", or "malicious people will just use this information to take advantage of me and interpret everything in the worst possible light" or "I'm not for public consumption, I am my own sovereign person, I don't owe everyone an explanation or justification for actions I have a right to do", then you'll want less optimization.

Of course, these aren't so much static "personality traits" of a person as one's assessment of the situation around oneself. The latter cluster is an assumption that you're living in a social environment where there's *very little concordance of interests*

— people knowing more about you will allow them to more effectively harm you. The former cluster is an assumption that you're living in an environment where there's a *great deal of concordance of interests* — people knowing more about you will allow them to more effectively help you.

For instance, being "predictable" is, in Venkat's writing, usually a bad thing, because it means you can be exploited by adversaries. Free people are "inscrutable." In other contexts, such as parenting, being predictable is a *good* thing, because you *want* your kids to have an easier time learning how to "work" the house rules. You and your kid are not, most of the time, wily adversaries outwitting each other; conflicts are more likely to come from too much confusion or inconsistently enforced boundaries. Relationship advice and management advice usually recommends making yourself *easier* for your partners and employees to understand, never more inscrutable. (Sales advice, however, and occasionally advice for keeping romance alive in a marriage, sometimes recommends cultivating an aura of mystery, perhaps because it's more adversarial.)

A related notion: wanting to join discussions is a sign of expecting a more cooperative world, while trying to keep people from joining your (private or illegible) communications is a sign of expecting a more adversarial world.

As social organizations "mature" and become larger, it becomes harder to enforce universal and impartial rules, harder to keep the larger population aligned on similar goals, and harder to comprehend the more complex phenomena in this larger group. . This means that there's both motivation and opportunity to carve out "hidden" and "special" zones where arbitrary behavior can persist even when it would otherwise come with negative consequences.

New or small organizations, by contrast, must gain/create resources or die, so they have more motivation to "optimize" for resource production; and they're simple, small, and/or homogeneous enough that legible optimization rules and goals and transparent communication are practical and widely embraced. "Security" is not available to begin with, so people mostly seek opportunity instead.

This theory explains, for instance, why US public policy is more fragmented, discretionary, and special-case-y, and less efficient and technocratic, than it is in other developed countries: the US is more racially diverse, which means, in a world where racism exists, that US civil institutions have evolved to allow ample opportunities to "play favorites" (giving special legal privileges to those with clout) *in full generality,* because a large population has historically been highly motivated to "play favorites" on the basis of race. Homogeneity makes a polity behave more like a "smaller" one, while diversity makes a polity behave more like a "larger" one.

Aesthetically, I think of optimization as corresponding to an "early" style, like Doric columns, or like Masaccio; simple, martial, all form and principle. Arbitrariness corresponds to a "late" style, like Corinthian columns or like Rubens: elaborate, sensual, full of details and personality.

The basic argument for optimization over arbitrariness is that it creates growth and value while arbitrariness creates stagnation.

Arbitrariness can't really argue for itself as well, because communication itself is on the other side. Arbitrariness always looks illogical and inconsistent. It kind of *is* illogical and inconsistent. All it can say is "I'm going to defend my right to be wrong, because I don't trust the world to understand me when I have a

counterintuitive or hard-to-express or controversial reason for my choice. I don't think I *can* get what I want by asking for it or explaining my reasons or playing 'fair'." And from the outside, you can't always tell the difference between someone who thinks (perhaps correctly!) that the game is really rigged against them a profound level, and somebody who just wants to cheat or who isn't thinking coherently. Sufficiently advanced cynicism is indistinguishable from malice *and* stupidity.

For a fairly sympathetic example, you see something like Darkness at Noon, where the protagonist thinks, "Logic inexorably points to Stalinism; but Stalinism is awful! Therefore, let me insist on some space free from the depredations of logic, some space where justice can be tempered by mercy and reason by emotion." From the distance of many years, it's easy to say that's silly, that of course there are *reasons* not to support Stalin's purges, that it's totally unnecessary to *reject logic and justice* in order to object to killing innocents. But from inside the system, if all the arguments you know how to formulate are Stalinist, if all the "shoulds" and "oughts" around you are Stalinist, perhaps all you can articulate at first is "I know all this is *right*, of course, but I don't *like* it."

Not everything people *call* reason, logic, justice, or optimization, *is* in fact reasonable, logical, just, or optimal; so, a person needs some defenses against those claims of superiority. In particular, defenses that can shelter them *even when they don't know what's wrong with the claims*. And that's the closest thing we get to an argument in favor of arbitrariness. It's actually not a bad point, in many contexts. The counterargument usually has to boil down to hope — to a sense of "I bet we can do better."

Literature Review: Distributed Teams

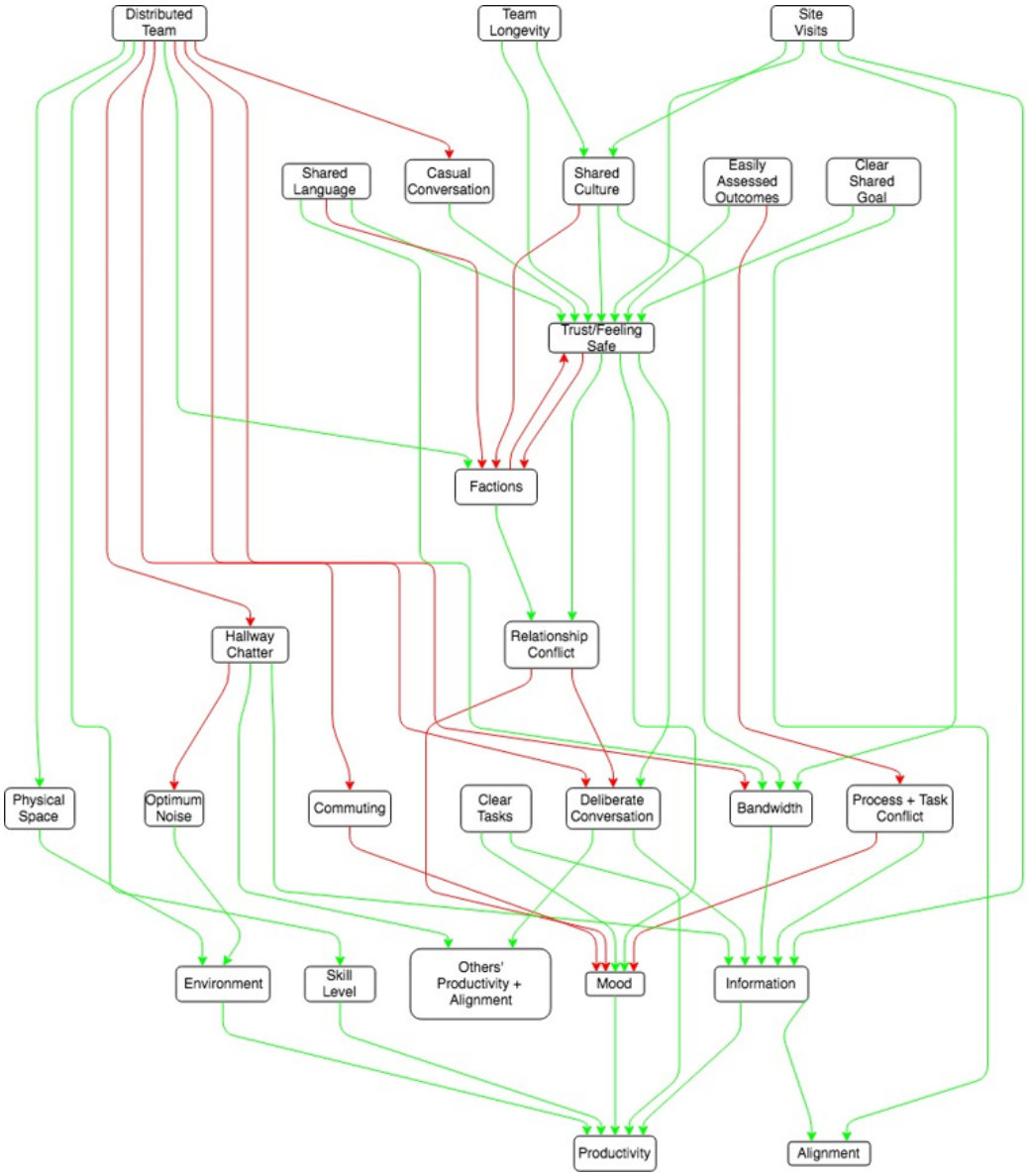
Introduction

Context: Oliver Habryka commissioned me to study and summarize the literature on distributed teams, with the goal of improving altruistic organizations. We wanted this to be rigorous as possible; unfortunately the rigor ceiling was low, for reasons discussed below. To fill in the gaps and especially to create a unified model instead of a series of isolated facts, I relied heavily on my own experience on a variety of team types (the favorite of which was an entirely remote company).

This document consists of five parts:

- Summary
- A series of specific questions Oliver asked, with supporting points and citations. My full, disorganized notes will be published as a comment.

My overall model of worker productivity is as follows:



Highlights and embellishments:

- Distribution decreases bandwidth and trust (although you can make up for a surprising amount of this with well timed visits).
- Semi-distributed teams are worse than fully remote or fully co-located teams on basically every metric. The politics are worse because geography becomes a fault line for factions, and information is lost because people incorrectly count on proximity to distribute information.
- You can get co-location benefits for about as many people as you can fit in a hallway: after that you're paying the costs of co-location while benefits decrease.
- No paper even attempted to examine the increase in worker quality/fit you can get from fully remote teams.

Sources of difficulty:

- Business science research is generally crap.
- Much of the research was quite old, and I expect technology to improve results from distribution every year.
- Numerical rigor trades off against nuance. This was especially detrimental when it comes to forming a model of how co-location affects politics, where much that happens is subtle and unseen. The most largest studies are generally survey data, which can only use crude correlations. The most interesting studies involved researchers reading all of a team's correspondence over months and conducting in-depth interviews, which can only be done for a handful of teams per paper.

How does distribution affect information flow?

"Co-location" can mean two things: actually working together side by side on the same task, or working in parallel on different tasks near each other. The former has an information bandwidth that technology cannot yet duplicate. The latter can lead to serendipitous information sharing, but also imposes costs in the form of noise pollution and siphoning brain power for social relations.

Distributed teams require information sharing processes to replace the serendipitous information sharing. These processes are less likely to be developed in teams with multiple locations (as opposed to entirely remote). Worst of all is being a lone remote worker on a co-located team; you will miss too much information and it's feasible only occasionally, despite the fact that measured productivity tends to rise when people work from home.

I think relying on co-location over processes for information sharing is similar to relying on human memory over writing things down: much cheaper until it hits a sharp cliff. Empirically that cliff is about 30 meters, or one hallway. After that, process shines.

List of isolated facts, with attribution:

- "The mutual knowledge problem" (Cramton 2015):
    - Assumption knowledge is shared when it is not, including:
        - typical minding.
        - Not realizing how big a request is (e.g. "why don't you just walk down the hall to check?", not realizing the lab with the data is 3 hours away. And the recipient of the request not knowing the asker does not know that, and so assumes the asker does not value their time).
    - Counting on informal information distribution mechanisms that don't distribute evenly
    - Silence can be mean many things and is often misinterpreted. E.g. acquiescence, deliberate snub, message never received.
- Lack of easy common language can be an incredible stressor and hamper information flow (Cramton 2015).
- People commonly cite overhearing hallway conversation as a benefit of co-location. My experience is that Slack is superior for producing this because it can be done asynchronously, but there's reason to believe I'm an outlier.
- Serendipitous discovery and collaboration falls off by the time you reach 30 meters (chapter 5), or once you're off the same hallway (chapter 6)

- Being near executives, project decision makers, sources of information (e.g. customers), or simply more of your peers gets you more information (Hinds, Retelny, and Cramton 2015)

How does distribution interact with conflict?

Distribution increases conflict and reduces trust in a variety of ways.

- Distribution doesn't lead to factions in and of itself, but can in the presence of other factors correlated with location
    - e.g. if the engineering team is in SF and the finance team in NY, that's two correlated traits for fault lines to form around. Conversely, having common traits across locations (e.g. work role, being parents of young children)] fights factionalization (Cramton and Hinds 2005).
    - Language is an especially likely fault line.
- Levels of trust and positive affect are generally lower among distributed teams (Mortenson and Neeley 2012) and even co-located people who work from home frequently enough (Gajendra and Harrison 2007).
- Conflict is generally higher in distributed teams (O'Leary and Mortenson 2009, Martins, Gilson, and Maynard 2004)
- It's easier for conflict to result in withdrawal among workers who aren't co-located, amplifying the costs and making problem solving harder.
- People are more likely to commit the fundamental attribution error against remote teammates (Wilson et al 2008).
- Different social norms or lack of information about colleagues lead to misinterpretation of behavior (Cramton 2016) e.g.,
    - you don't realize your remote co-worker never smiles at anyone and so assume he hates you personally.
    - different ideas of the meaning of words like "yes" or "deadline".
- From analogy to biology I predict conflict is most likely to arise when two teams are relatively evenly matched in terms of power/ resources and when spoils are winner take all.
- Most site:site conflict is ultimately driven by desire for access to growth opportunities (Hinds, Retelny, and Cramton 2015). It's not clear to me this would go away if everyone is co-located- it's easier to view a distant colleague as a threat than a close one, but if the number of opportunities is the same, moving people closer doesn't make them not threats.
- Note that conflict is not always bad- it can mean people are honing their ideas against others'. However the literature on virtual teams is implicitly talking about relationship conflict, which tends to be a pure negative.

When are remote teams preferable?

- You need more people than can fit in a 30m radius circle (chapter 5), or a single hallway. (chapter 6).
- Multiple critical people can't be co-located, e.g.,
    - Wave's compliance officer wouldn't leave semi-rural Pennsylvania, and there was no way to get a good team assembled there.

    - Lobbying must be based in Washington, manufacturing must be based somewhere cheaper.
    - Customers are located in multiple locations, such that you can co-locate with your team members or customers, but not both.
- If you must have some team members not co-located, better to be entirely remote than leave them isolated. If most of the team is co-located, they will not do the things necessary to keep remote individuals in the loop.
- There is a clear shared goal
- The team will be working together for a long time and knows it (Alge, Weithoff, and Klein 2003)
- Tasks are separable and independent.
- You can filter for people who are good at remote work (independent, good at learning from written work).
- The work is easy to evaluate based on outcome or produces highly visible artifacts.
- The work or worker benefits from being done intermittently, or doesn't lend itself to 8-hours-and-done, e.g.,
    - Wave's anti-fraud officer worked when the suspected fraud was happening.
    - Engineer on call shifts.
- You need to be process- or documentation-heavy for other reasons, e.g. legal, or find it relatively cheap to be so (chapter 2).
- You want to reduce variation in how much people contribute (=get shy people to talk more) (Martins, Gilson, and Maynard 2008).
- Your work benefits from long OODA loops.
- You anticipate low turnover (chapter 2).

How to mitigate the costs of distribution

- Site visits and retreats, especially early in the process and at critical decision points. I don't trust the papers quantitatively, but some report site visits doing as good a job at trust- and rapport-building as co-location, so it's probably at least that order of magnitude (see Hinds and Cramton 2014 for a long list of studies showing good results from site visits).
    - Site visits should include social activities and meals, not just work. Having someone visit and not integrating them socially is worse than no visit at all.
    - Site visits are more helpful than retreats because they give the visitor more context about their coworkers (chapter 2). This probably applies more strongly in industrial settings.
- Use voice or video when need for bandwidth is higher (chapter 2).
    - Although high-bandwidth virtual communication may make it easier to lie or mislead than either in person or low-bandwidth virtual communication (Håkonsson *et al* 2016).
- Make people very accessible, e.g.,
    - Wave asked that all employees leave skype on autoanswer while working, to recreate walking to someone's desk and tapping them on the shoulder.
    - Put contact information in an accessible wiki or on Slack, instead of making people ask for it.
- Lightweight channels for building rapport, e.g., CEA's compliments Slack channel, Wave's kudos section in weekly meeting minutes (personal

observation).
- Build over-communication into the process.
    - In particular, don't let silence carry information. Silence can be interpreted a million different ways (Cramton 2001).
- Things that are good all the time but become more critical on remote teams
    - Clear goals/objectives
    - Clear metrics for your goals/objectives
    - Clear roles (Zacarro, Ardison, Orvis 2004)
    - Regular 1:1s
    - Clear communication around current status
    - Long time horizons (chapter 10).
    - Shared identity (Hinds and Mortensen 2005) with identifiers (chapter 10), e.g. t-shirts with logos.
- Have a common chat tool (e.g., Slack or Discord) and give workers access to as many channels as you can, to recreate hallway serendipity (personal observation).
- Hire people like me
    - long OODA loop
    - good at learning from written information
    - Good at working working asynchronously
    - Don't require social stimulation from work
- Be fully remote, as opposed to just a few people working remotely or multiple co-location sites.
- If you have multiple sites, lumping together similar people or functions will lead to more factions (Cramton and Hinds 2005). But co-locating people who need to work together takes advantage of the higher bandwidth co-location provides..
- Train workers in active listening (chapter 4) and conflict resolution. Microsoft uses the *Crucial Conversations* class, and I found the book of the same name incredibly helpful.

Cramton 2016 was an excellent summary paper I refer to a lot in this write up. It's not easily available on-line, but the author was kind enough to share a PDF with me that I can pass on.

My full notes will be published as a comment on this post.

How do people become ambitious?

I have a pet theory about how people become ambitious and agenty.

I was about to write up some insight porn about it, and then was like "you know, Raemon, you should probably actually think about about this for real, since it seems like Pet Psychology Theories are one of the easier ways to get stuck in dumb cognitive traps."

"How do people get ambitious and agenty" seems to be something people should have actually tried to study before. I'm thinking something as simple as "interviewing lots of people and checking for common patterns."

2 seconds spent on google scholar suggested I could use better keywords.

Curious if anyone has looked into this (either reviewing existing literature, or conducting interviews themselves or otherwise trying to tackle the question in a serious way)

For clarity, the two phenomena I want to understand better are:

- *Ambition.* How do people end up having plans (which they realistically expect to achieve) that affect at least thousands, and preferably millions of people?
- *Agency.* How do people generally gain the capacity to be self-motivated, think through plans, and decide how to pursue goals that will change the world (changing it in small ways is fine)

1960: The Year The Singularity Was Cancelled

*[**Epistemic status:** Very speculative, especially Parts 3 and 4. Like many good things, this post is based on a conversation with Paul Christiano; most of the good ideas are his, any errors are mine.]*

I.

In the 1950s, an Austrian scientist discovered a series of equations that he claimed could model history. They matched past data with startling accuracy. But when extended into the future, they predicted the world would end on November 13, 2026.

This sounds like the plot of a sci-fi book. But it's also the story of Heinz von Foerster, a mid-century physicist, cybernetician, cognitive scientist, and philosopher.

His problems started when he became interested in human population dynamics.

(the rest of this section is loosely adapted from his *Science* paper "Doomsday: Friday, 13 November, A.D. 2026")

Assume a perfect paradisiacal Garden of Eden with infinite resources. Start with two people – Adam and Eve – and assume the population doubles every generation. In the second generation there are 4 people; in the third, 8. This is that old riddle about the grains of rice on the chessboard again. By the 64th generation (ie after about 1500 years) there will be 18,446,744,073,709,551,615 people – ie about about a billion times the number of people who have ever lived in all the eons of human history. So one of our assumptions must be wrong. Probably it's the one about the perfect paradise with unlimited resources.

Okay, new plan. Assume a limited world with a limited food supply / limited carrying capacity. If you want, imagine it as an island where everyone eats coconuts. But there are only enough coconuts to support 100 people. If the population reproduces beyond 100 people, some of them will starve, until they're back at 100 people. In the second generation, there are 100 people. In the third generation, still 100 people. And so on to infinity. Here the population never grows at all. But that doesn't match real life either.

But von Foerster knew that technological advance can change the carrying capacity of an area of land. If our hypothetical islanders discover new coconut-tree-farming techniques, they may be able to get twice as much food, increasing the maximum population to 200. If they learn to fish, they might open up entirely new realms of food production, increasing population into the thousands.

So the rate of population growth is neither the double-per-generation of a perfect paradise, nor the zero-per-generation of a stagnant island. Rather, it depends on the rate of economic and technological growth. In particular, in a closed system that is already at its carrying capacity and with zero marginal return to extra labor, population growth equals productivity growth.

What causes productivity growth? Technological advance. What causes technological advance? Lots of things, but von Foerster's model reduced it to one: people. Each person has a certain percent chance of coming up with a new discovery that improves the economy, so productivity growth will be a function of population.

So in the model, the first generation will come up with some small number of technological advances. This allows them to spawn a slightly bigger second generation. This new slightly larger population will generate slightly more technological advances. So each generation, the population will grow at a slightly faster rate than the generation before.

This matches reality. The world population barely increased at all in the millennium from 2000 BC to 1000 BC. But it doubled in the fifty years from 1910 to 1960. In fact, using his model, von Foerster was able to come up with an equation that predicted the population near-perfectly from the Stone Age until his own day.

But his equations corresponded to something called hyperbolic growth. In hyperbolic growth, a feedback cycle – in this case population causes technology causes more population causes more technology – leads to growth increasing rapidly and finally shooting to infinity. Imagine a simplified version of Foerster's system where the world starts with 100 million people in 1 AD and a doubling time of 1000 years, and the doubling time decreases by half after each doubling. It might predict something like this:

1 AD: 100 million people
1000 AD: 200 million people
1500 AD: 400 million people
1750 AD: 800 million people
1875 AD: 1600 million people

…and so on. This system reaches infinite population in finite time (ie before the year 2000). The real model that von Foerster got after analyzing real population growth was pretty similar to this, except that it reached infinite population in 2026, give or take a few years (his pinpointing of Friday November 13 was mostly a joke; the equations were not really that precise).

What went wrong? Two things.

First, as von Foerster knew (again, it was kind of a joke) the technological advance model isn't literally true. His hyperbolic model just operates as an upper bound on the Garden of Eden scenario. Even in the Garden of Eden, population can't do more than double every generation.

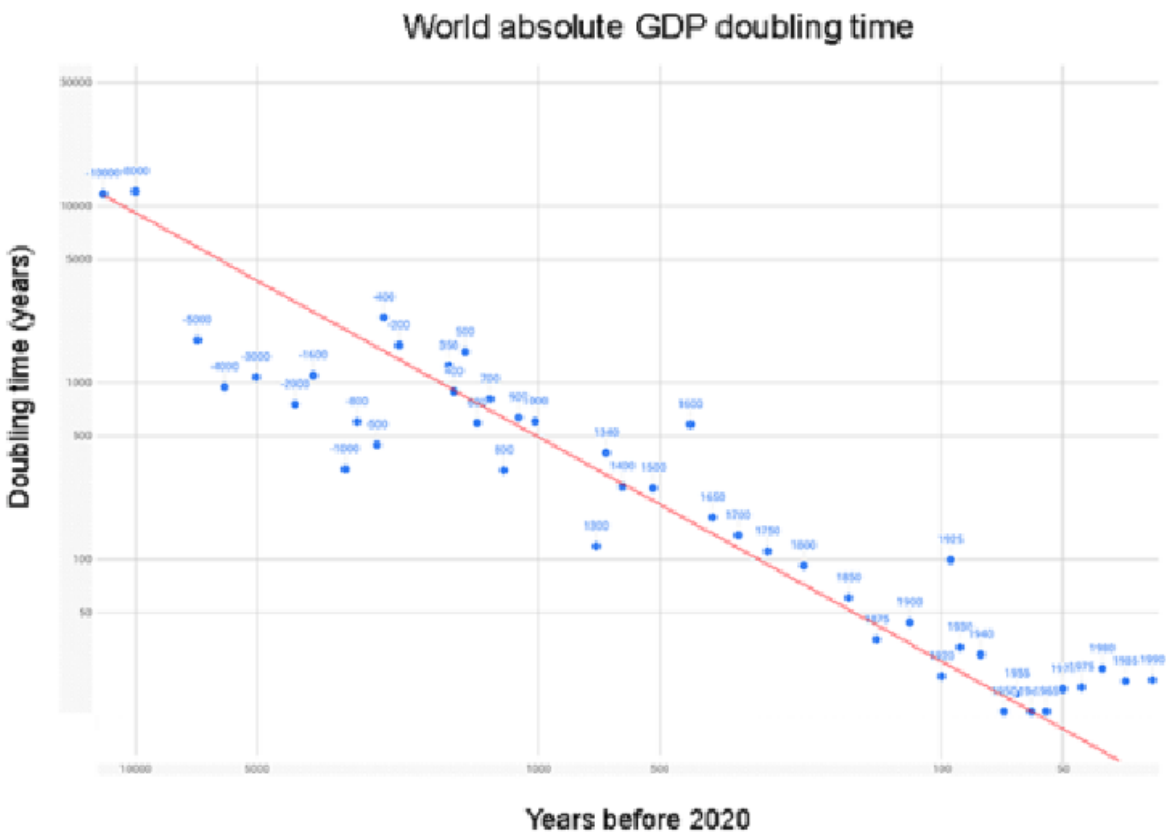
Second, contra all previous history, people in the 1900s started to have fewer kids than their resources could support (the demographic transition). Couples started considering the cost of college, and the difficulty of maternity leave, and all that, and decided that maybe they should stop at 2.5 kids (or just get a puppy instead).

Von Foerster published has paper in 1960, which ironically was the last year that his equations held true. Starting in 1961, population left its hyperbolic growth path. It is now expected to stabilize by the end of the 21st century.

II.

But nobody really expected the population to reach infinity. Armed with this story, let's look at something more interesting.

This might be the most depressing graph ever:



The horizontal axis is years before 2020, a random year chosen so that we can put this in log scale without negative values screwing everything up.

The vertical axis is the amount of time it took the world economy to double from that year, according to this paper. So for example, if at some point the economy doubled every twenty years, the dot for that point is at twenty. The doubling time decreases throughout most of the period being examined, indicating hyperbolic growth.

Hyperbolic growth, as mentioned before, shoots to infinity at some specific point. On this graph, that point is represented by the doubling time reaching zero. Once the economy doubles every zero years, you might as well call it infinite.

For all of human history, economic progress formed a near-perfect straight line pointed at the early 21st century. Its destination varied by a century or two now and then, but never more than that. If an ancient Egyptian economist had modern techniques and methodologies, he could have made a graph like this and predicted it would reach infinity around the early 21st century. If a Roman had done the same thing, using the economic data available in his own time, he would have predicted the early 21st century too. A medieval Burugundian? Early 21st century. A Victorian Englishman? Early 21st century. A Stalinist Russian? Early 21st century. The trend was *really* resilient.

In 2005, inventor Ray Kurzweil published *The Singularity Is Near*, claiming there would be a technological singularity in the early 21st century. He didn't refer to this graph specifically, but he highlighted this same trend of everything getting faster, including rates of change. Kurzweil took the infinity at the end of this graph very seriously; he

thought that some event would happen that really *would* catapult the economy to infinity. Why not? Every data point from the Stone Age to the Atomic Age agreed on this.

This graph shows the Singularity getting cancelled.

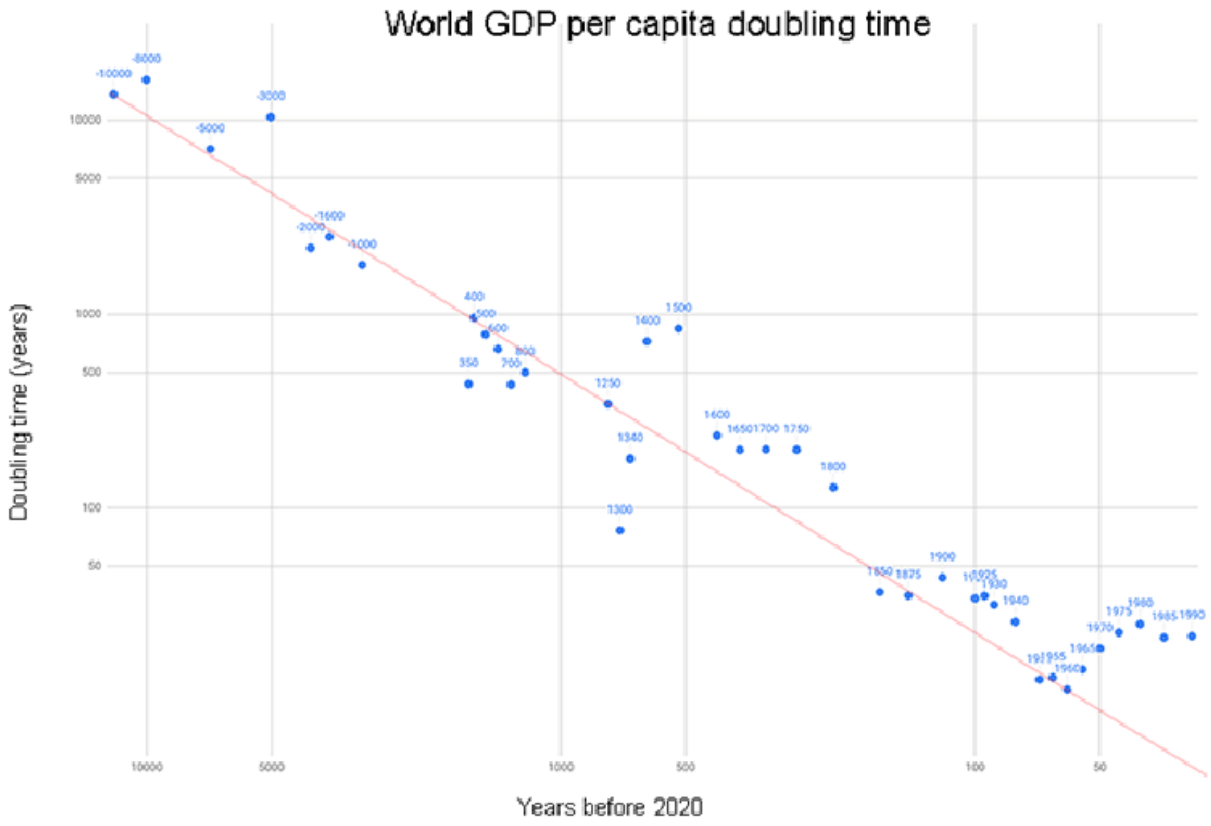
Around 1960, doubling times stopped decreasing. The economy kept growing. But now it grows at a flat rate. It shows no signs of reaching infinity; not soon, not ever. Just constant, boring 2% GDP growth for the rest of time.

Why?

Here von Foerster has a ready answer prepared for us: population!

Economic growth is a function of population and productivity. And productivity depends on technological advancement and technological advancement depends on population, so it all bottoms out in population in the end. And population looked like it was going to grow hyperbolically until 1960, after which it stopped. That's why hyperbolic economic growth, ie progress towards an economic singularity, stopped then too.

In fact...



This is a *really sketchy* of per capita income doubling times. It's sketchy because until 1650, per capita income wasn't really increasing at all. It was following a one-step-forward one-step-back pattern. But if you take out all the steps back and just watch how quickly it took the steps forward, you get something like this.

Even though per capita income tries to abstract out population, it displays the same pattern. Until 1960, we were on track for a singularity where everyone earned infinite money. After 1960, the graph “bounces back” and growth rates stabilize or even decrease.

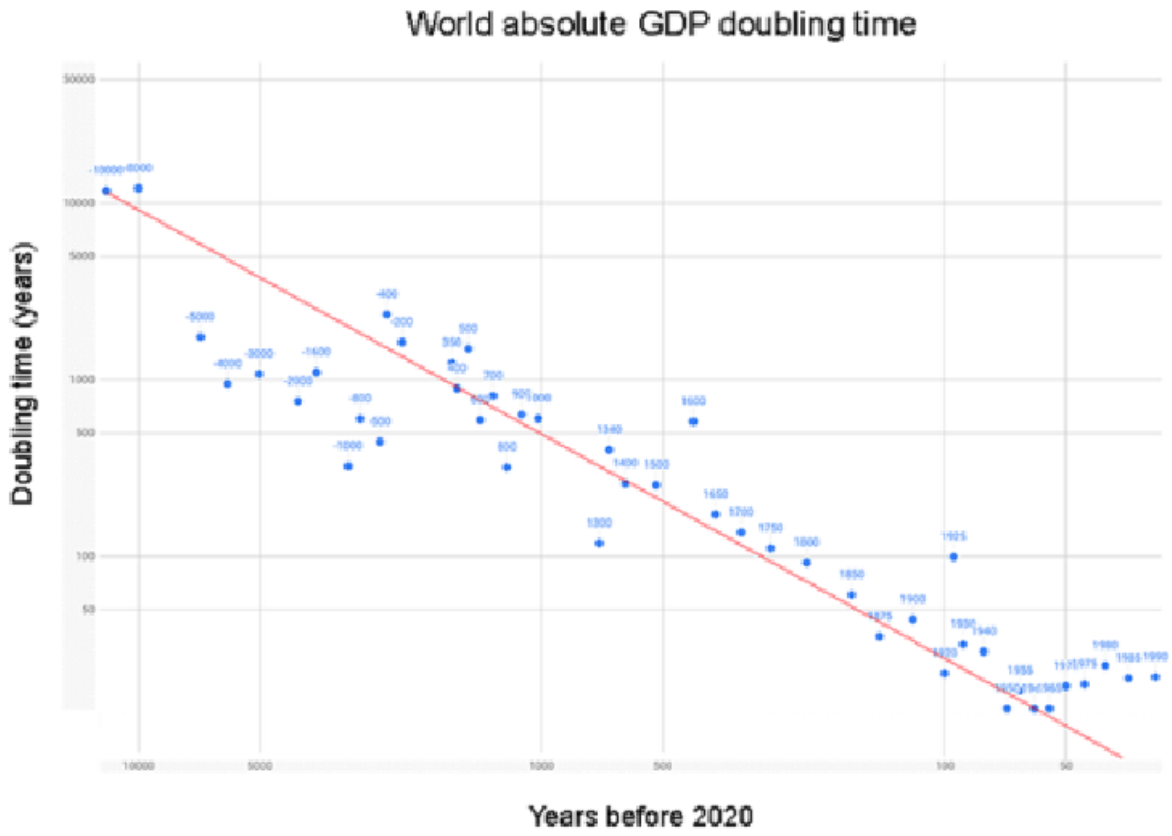
Again, von Foerster can explain this to us. Per capita income grows when technology grows, and technology grows when the population grows. The signal from the end of hyperbolic population growth shows up here too.

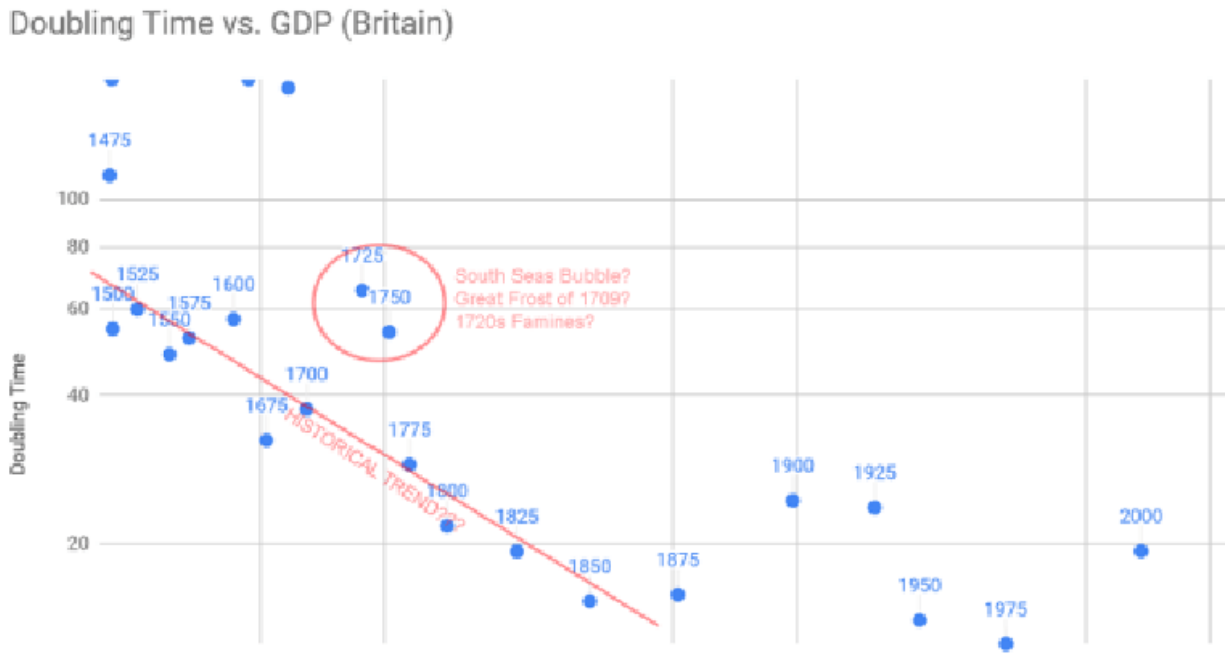
To make this really work, we probably have to zoom in a little bit and look at concrete reality. Most technological advances come from a few advanced countries whose population stabilized a little earlier than the world population. Of the constant population, an increasing fraction are becoming researchers each year (on the other hand, the low-hanging fruit gets picked off and technological advance becomes harder with time). All of these factors mean we shouldn’t expect productivity growth/GWP per capita growth/technological growth to *exactly* track population growth. But on the sort of orders-of-magnitude scale you can see on logarithmic graphs like the ones above, it should be pretty close.

So it looks like past predictions of a techno-economic singularity for the early 21st century were based on extrapolations of a hyperbolic trend in technology/economy that depended on a hyperbolic trend in population. Since the population singularity didn’t pan out, we shouldn’t expect the techno-economic singularity to pan out either. In fact, since population in advanced countries is starting to “stagnate” relative to earlier eras, we should expect a relative techno-economic stagnation too.

…maybe. Before coming back to this, let’s explore some of the other implications of these models.

III.

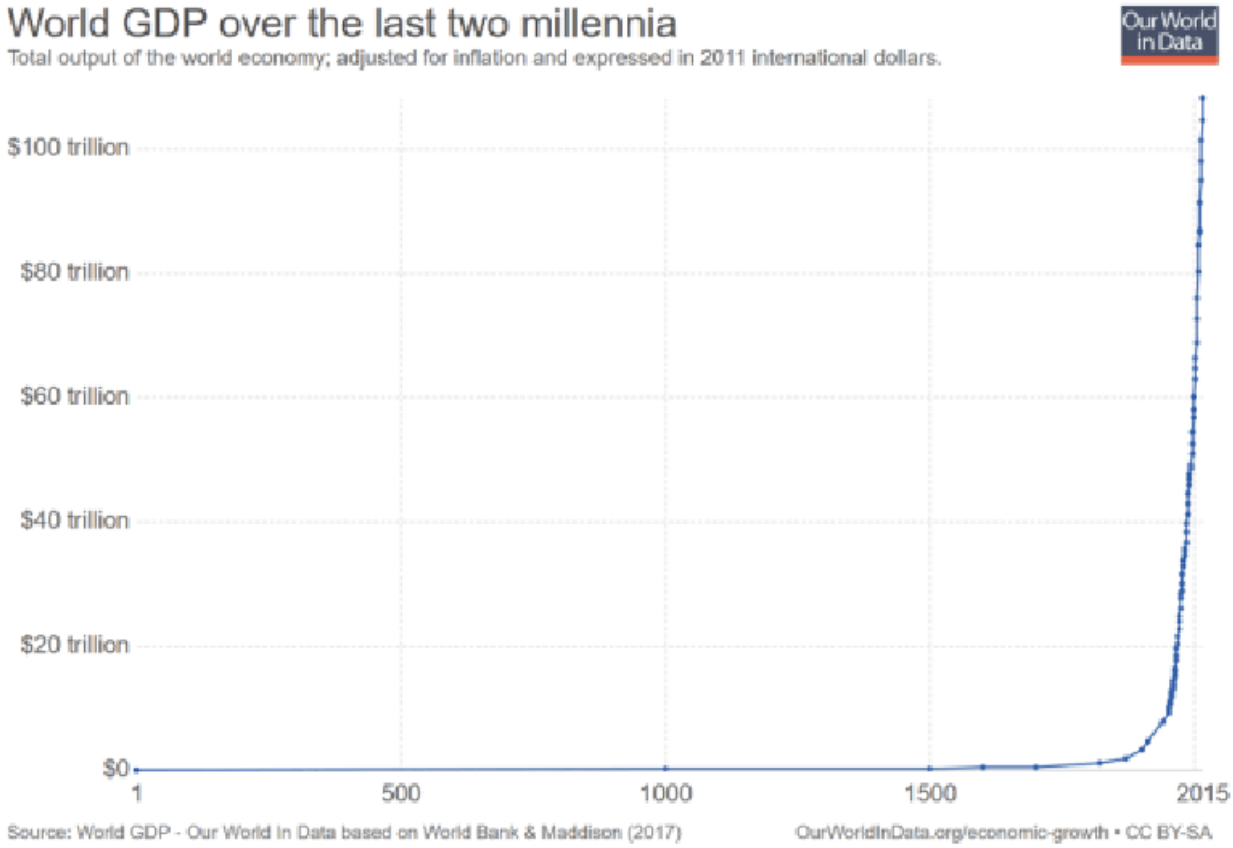




The first graph is the same one you saw in the last section, of absolute GWP doubling times. The second graph is the same, but limited to Britain.

Where’s the Industrial Revolution?

It doesn’t show up at all. This may be a surprise if you’re used to the standard narrative where the Industrial Revolution was the most important event in economic history. Graphs like this make the case that the Industrial Revolution was an explosive shift to a totally new growth regime:



It sure *looks* like the Industrial Revolution was a big deal. But Paul Christiano argues your eyes may be deceiving you. That graph is a hyperbola, ie corresponds to a single simple equation. There is no break in the pattern at any point. If you transformed it to a log doubling time graph, you’d just get the graph above that looks like a straight line until 1960.

On this view, the Industiral Revolution didn’t change historical GDP trends. It just shifted the world from a Malthusian regime where economic growth increased the population to a modern regime where economic growth increased per capita income.

For the entire history of the world until 1000, GDP per capita was the same for everyone everywhere during all historical eras. An Israelite shepherd would have had about as much stuff as a Roman farmer or a medieval serf.

This was the Malthusian trap, where “productivity produces people, not prosperity”. People reproduce to fill the resources available to them. Everyone always lives at subsistence level. If productivity increases, people reproduce, and now you have more people living at subsistence level. OurWorldInData has an awesome graph of this:

As of 1500, places with higher productivity (usually richer farmland, but better technology and social organization also help) population density is higher. But GDP per capita was about the same everywhere.

There were always occasional windfalls from exciting discoveries or economic reforms. For a century or two, GDP per capita would rise. But population would always catch up again, and everyone would end up back at subsistence.

Some people argue Europe broke out of the Malthusian trap around 1300. This is not quite right. 1300s Europe achieved above-subsistence GDP, but only because the Black Plague killed so many people that the survivors got a windfall by taking their land.

Malthus predicts that this should only last a little while, until the European population bounces back to pre-Plague levels. This prediction was exactly right for Southern Europe. Northern Europe didn't bounce back. Why not?

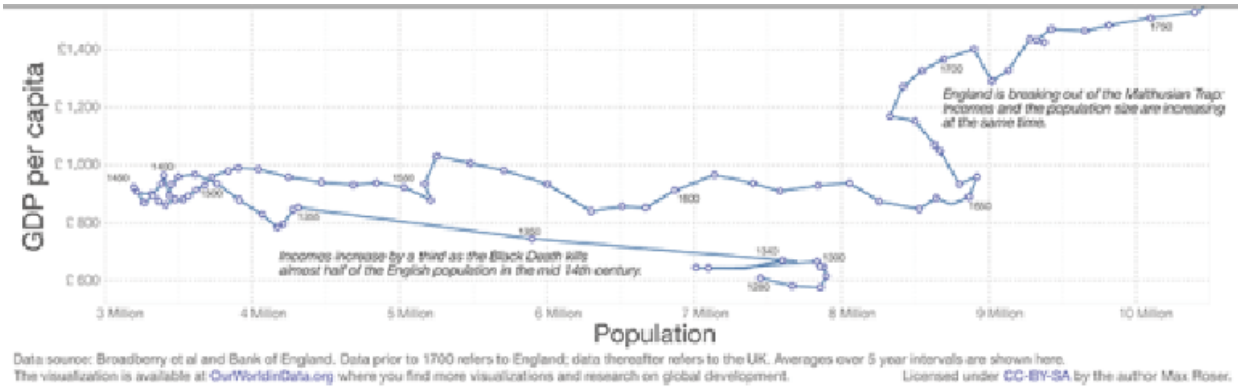
Unclear, but one answer is: fewer people, more plagues.

Broadberry 2015 mentions that Northern European culture promoted later marriage and fewer children:

> The North Sea Area had an advantage in this area because of its approach to marriage. Hajnal (1965) argued that northwest Europe had a different demographic regime from the rest of the world, characterised by later marriage and hence limited fertility. Although he originally called this the European Marriage Pattern, later work established that it applied only to the northwest of the continent. This can be linked to the availability of labour market opportunities for females, who could engage in market activity before marriage, thus increasing the age of first marriage for females and reducing the number of children conceived (de Moor and van Zanden, 2010). Later marriage and fewer children are associated with more investment in human capital, since the womenemployed in productive work can

accumulate skills, and parents can afford to invest more in each of the smaller number of children because of the “quantity-quality” trade-off (Voigtländer and Voth, 2010).

This low birth rate was happening at the same time plagues were raising the death rate. Here’s another amazing graph from OurWorldInData:



British population maxes out around 1300 (?), declines substantially during the Black Plague of 1348-49, but then keeps declining. The List Of English Plagues says another plague hit in 1361, then another in 1369, then another in 1375, and so on. Some historians call the whole period from 1348 to 1666 “the Plague Years”.

It looks like through the 1350 – 1450 period, population keeps declining, and per capita income keeps going up, as Malthusian theory would predict.

Between 1450 and 1550, population starts to recover, and per capita incomes start going down, again as Malthus would predict. Then around 1560, there’s a jump in incomes; according to the List Of Plagues, 1563 was “probably the worst of the great metropolitan epidemics, and then extended as a major national outbreak”. After 1563, population increases again and per capita incomes decline again, all the way until 1650. Population does not increase in Britain at all between 1660 and 1700. Why? The List declares 1665 to be “The Great Plague”, the largest in England since 1348.

So from 1348 to 1650, Northern European per capita incomes diverged from the rest of the world’s. But they didn’t “break out of the Malthusian trap” in a strict sense of being able to direct production toward prosperity rather than population growth. They just had so many plagues that they couldn’t grow the population anyway.

But in 1650, England did start breaking out of the Malthusian trap; population and per capita incomes grow together. Why?

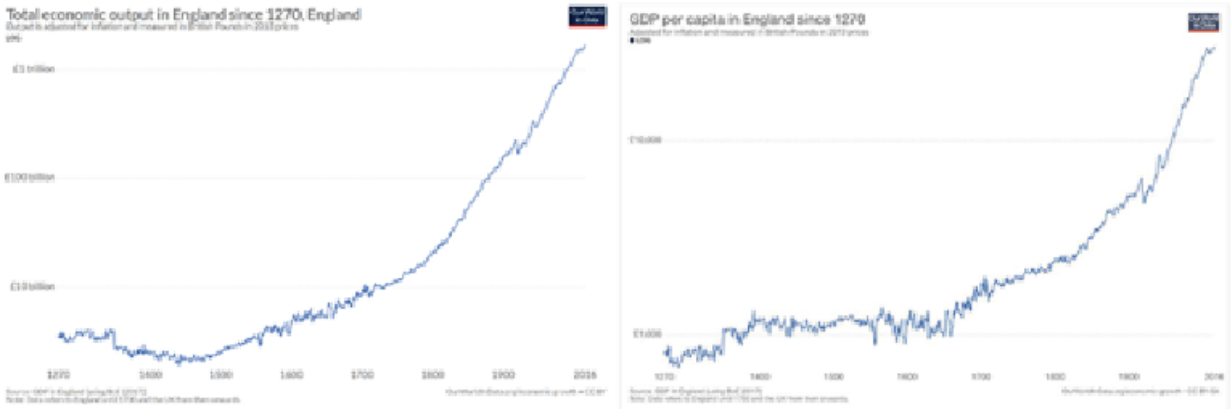
Paul theorizes that technological advance finally started moving faster than maximal population growth.

Remember, in the von Foerster model, the growth rate increases with time, all the way until it reaches infinity in 2026. The closer you are to 2026, the faster your economy will grow. But population can only grow at a limited rate. In the absolute limit, women can only have one child per nine months. In reality, infant mortality, infertility, and conscious decision to delay childbearing mean the natural limits are much lower than that. So there’s a theoretical limit on how quickly the population can increase even with maximal resources. If the economy is growing faster than that, Malthus can’t catch up.

Why would this happen in England and Holland in 1650?

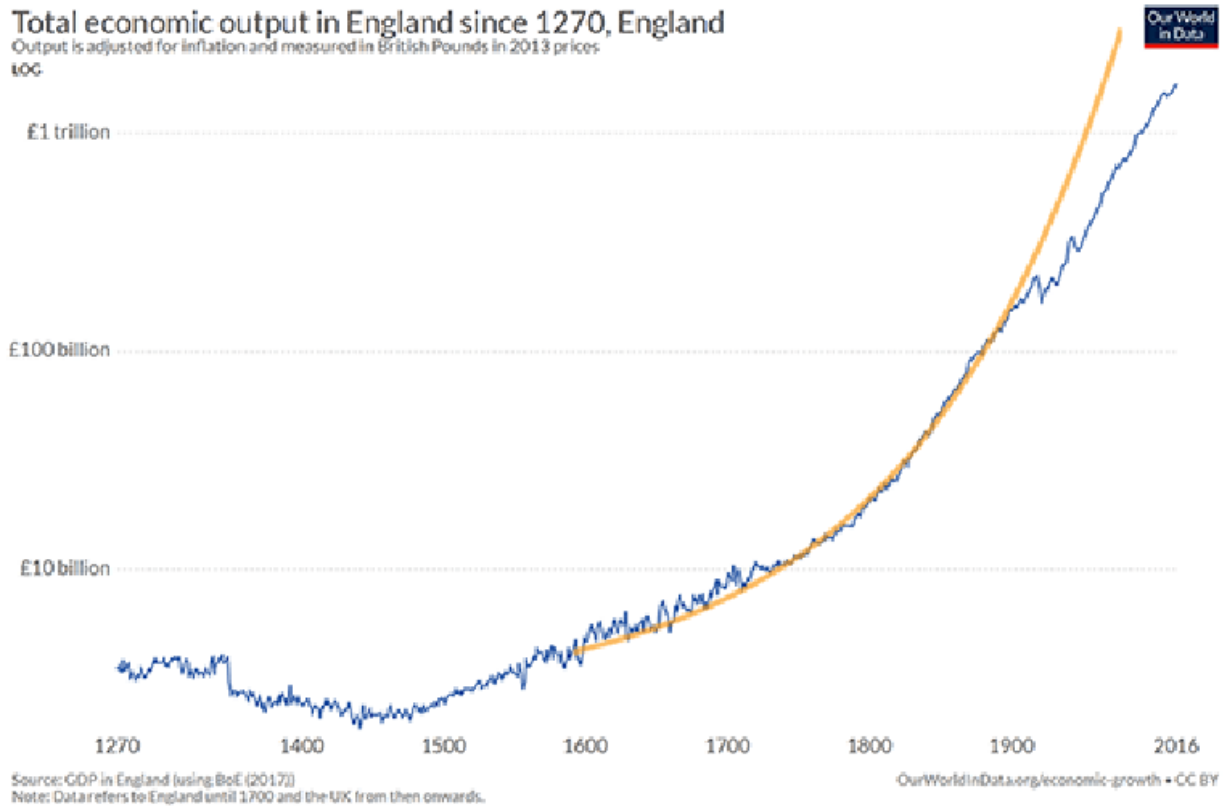
Lots of people have historical explanations for this. Northern European population growth was so low that people were forced to invent labor-saving machinery; eventually this reached a critical mass, we got the Industrial Revolution, and economic growth skyrocketed. Or: the discovery of America led to a source of new riches and a convenient sink for excess population. Or: something something Protestant work ethic printing press capitalism. These are all plausible. But how do they sync with the claim that absolute GDP never left its expected trajectory?

I find the idea that the Industrial Revolution wasn't a deviation from trend fascinating and provocative. But it depends on eyeballing a lot of graphs that have had a lot of weird transformations done to them, plus writing off a lot of outliers. Here's another way of presenting Britain's GDP and GDP per capita data:



Here it's a lot less obvious that the Industrial Revolution represented a deviation from trend for GDP per capita but not for GDP.

These British graphs show less of a singularity signature than the worldwide graphs do, probably because we're looking at them on a shorter timeline, and because the Plague Years screwed everything up. If we insisted on fitting them to a hyperbola, it would look like this:



Like the rest of the world, Britain was only on a hyperbolic growth trajectory when economic growth was translating into population growth. That wasn't true before about 1650, because of the plague. And it wasn't true after about 1850, because of the Demographic Transition. We see a sort of fit to a hyperbola between those points, and then the trend just sort of wanders off.

It seems possible that the Industrial Revolution was not a time of abnormally fast technological advance or economic growth. Rather, it was a time when economic growth outpaced population growth, causing a shift from a Malthusian regime where productivity growth always increased population at subsistence level, to a modern regime where productivity growth increases GDP per capita. The world remained on the same hyperbolic growth trajectory throughout, until the trajectory petered out around 1900 in Britain and around 1960 in the world as a whole.

IV.

So just how cancelled is the singularity?

To review: population growth increases technological growth, which feeds back into the population growth rate in a cycle that reaches infinity in finite time.

But since population can't grow infinitely fast, this pattern breaks off after a while.

The Industrial Revolution tried hard to compensate for the "missing" population; it invented machines. Using machines, an individual could do an increasing amount of work. We can imagine making eg tractors as an attempt to increase the effective population faster than the human uterus can manage. It partly worked.

But the industrial growth mode had one major disadvantage over the Malthusian mode: tractors can't invent things. The population wasn't just there to grow the population, it was there to increase the rate of technological advance and thus population growth. When we shifted (in part) from making people to making tractors, that process broke down, and growth (in people *and* tractors) became sub-hyperbolic.

If the population stays the same (and by "the same", I just mean "not growing hyperbolically") we should expect the growth rate to stay the same too, instead of increasing the way it did for thousands of years of increasing population, modulo other concerns.

In other words, the singularity got cancelled because we no longer have a surefire way to convert money into researchers. The old way was more money = more food = more population = more researchers. The new way is just more money = send more people to college, and screw all that.

But AI potentially offers a way to convert money into researchers. Money = build more AIs = more research.

If this were true, then once AI comes around – even if it isn't much smarter than humans – then as long as the computational power you can invest into researching a given field increases with the amount of money you have, hyperbolic growth is back on. Faster growth rates means more money means more AIs researching new technology means even faster growth rates, and so on to infinity.

Presumably you would eventually hit some other bottleneck, but things could get very strange before that happens.

Helen Toner on China, CSET, and AI

This is a linkpost for http://rationallyspeakingpodcast.org/show/rs-231-helen-toner-on-misconceptions-about-china-and-artific.html

Miscellaneous excerpts I found interesting from Julia Galef's recent interview of Helen Toner:

Helen Toner: [...] [8:00] I feel like the West has this portrayal of the company Baidu as the "Google" of China. And when you think “Google,” you think super high tech, enormous, has many products that are really cool and really widely used. Like Google has search, obviously, but it also has Gmail. It also has Google Maps. It has a whole bunch of other things, Android...

So I feel like this term, “the Google of China,” gets applied to Baidu in all kinds of ways. And in fact, it's sort of true that Baidu is the main search engine in China, because Google withdrew from China. But kind of all the other associations we have with Google don't fit super well into Baidu. Maybe other than, it is one of China's largest tech companies, that is true.

But just in terms of my overall level of how impressed I was with Baidu as a company, or how much I expected them to do cool stuff in the future -- that went down by a lot, just based on... there's no... Baidu Maps exists, but no one really uses it. The most commonly used maps app is a totally different company. There's no Baidu Mail. There's no Baidu Docs. There's a lot of stories of management dysfunction or feudalism internally. So, that was one of the clearest updates I made.

Julia Galef: [...] [19:40] In your conversations in particular with the AI scientists who you got to meet in China, what did you notice? Did anything surprise you? Were their views different in any systematic way from the American AI scientists you'd talked to?

Helen Toner: Yeah. So, I should definitely caveat that this was a small number of conversations. It was maybe five conversations of any decent length.

Julia Galef: Oh, you also went to at least one AI conference, I know, in China.

Helen Toner: Yes, that's true. But it was much more difficult to have sort of substantive, in-depth conversations there.

I think a thing that I noticed in general among these conversations with more technical people... In the West, in similar conversations that I've been a part of, there's often been a part of the conversation that was dedicated to, "How do you think AI will affect society? What do you think are the important potential risks or benefits or whatever?" And maybe I have my own views, and I share those views. And usually the person doesn't 100% agree with me, and maybe they'll sort of provide a slightly different take or a totally different take, but they usually seem to have a reasonably well-thought-through picture of “What does AI mean for society? What might be good or bad about it?”

The Chinese professors that I talked to -- and this could totally just be a matter of relationships, that they didn't feel comfortable with me -- but they really didn't seem

interested in engaging in that part of the conversation. They sort of seemed to want to say things like, "Oh, it's just going to be a really good tool. So it will just do what humanity -- or, just do what its users will want it to do. And that's sort of all." And I would kind of ask about risks, and they would say, "Oh, that's not really something that I've thought about."

There's sort of an easy story you could tell there, which might be correct, which is basically: Chinese people are taught from a very young age that they should not have, or that it's dangerous to have, strong opinions about how the world should be and how society should be, and that the important thing is just to fall in line and do your job. So that's one possibility for what's going on. Of course, I might have just had selection bias, or they might have thought that I was this strange foreigner asking them strange questions and they didn't want to say anything. Who knows?

Julia Galef: Well, I mean, another possible story might be that the sources of the discourse around AI risk in the West just haven't permeated China. Like there's this whole discourse that got signal boosted with Elon Musk and so on. So, there have been all these conversations in our part of the world that just maybe aren't happening there.

Helen Toner: Sure, but I feel like plenty of the conversations I'm thinking of in the West happened before that was so widespread, and often the pushback would be something along the lines of, "Oh, no, those kinds of worries are not reasonable -- but I am really worried about employment, and here's how I think it's going to affect employment," or things along those lines. And that just didn't come up in any of these conversations, which I found a little bit surprising.

Julia Galef: Sure. Okay, so you got back from China recently. You became the director of strategy for CSET, the Center for Security and Emerging Technology. Can you tell us a little bit about why CSET was founded and what you're currently working on?

Helen Toner: Yeah. So, we were basically founded because -- so our name, the Center for Security and Emerging Technology, gives us some ability to be broad in what kinds of emerging technologies we focus on. For the first at least two years, we're planning to focus on AI and advanced computing. And that may well end up being more than two years, depending on how things play out. And so the reason we were founded is essentially because of seeing this gap in the supply and demand in DC, or the appetite, for analysis and information on how the US government should be thinking about AI, in all kinds of different ways. And so the one that we wanted to focus in on was the security or national security dimensions of that, because we think that they're so important, and we think that missteps there could be really damaging. So that's the basic overview of CSET.

Julia Galef: So it sounds like the reason that you decided to focus on AI specifically out of all possible emerging technologies is just because the supply and demand gap was especially large there for information?

Helen Toner: That's right. That's right.

So, what we work on in the future will similarly be determined by that. So, certainly on a scale of 10 or 20 years, I wouldn't want to be starting an organization that was definitely going to be working on AI for that length of time. So, depending on how things play out, we have room to move into different technologies where the government could use more in-depth analysis than it has time or resources to pursue.

Julia Galef: Great. And when you talk about AI, are you more interested in specialized AI -- like the kinds of things that are already in progress, like deep fakes or drones? Or are you more interested in the longer-term potential for general superintelligence?

Helen Toner: Yeah, so our big input into what we work on is what policymakers and other decision-makers in the government would find most useful. So that kind of necessarily means that we focus a lot on technologies that are currently in play or might be in play in sort of the foreseeable future. More speculative technologies can certainly come into our work if we think that that's relevant or important, but it's not our bread and butter. [...]

Julia Galef: [...] [26:15] In your interactions so far with American policymakers about AI, has anything surprised you about their views? Have there been any key disagreements that you find you have with the US policy community?

Helen Toner: I mean, I think an interesting thing about being in DC is just that everyone here, or so many people here, especially people in government, have so little time to think about so many issues. And there's so much going on, and they have to try and keep their heads wrapped around it.

So this means that kind of inevitably, simple versions of important ideas can be very powerful and can get really stuck in people's minds. So I see a few of these that I kind of disagree with, and I kind of understand why they got an embedded -- but if I had my druthers, I would embed a slightly different idea.

An example of this would be, in terms of AI, the idea that data is this super, super important input to machine learning systems. That's sort of step one of the argument, and step two of the argument is: And China has a larger population and weaker privacy controls, so Chinese companies and Chinese government will have access to more data. That’s step two. Therefore, conclusion: China has this intrinsic advantage in AI.

Julia Galef: Right. Yeah. I've heard that framed in terms of a metaphor where data is like oil.

Helen Toner: Right. Exactly.

Julia Galef: So, China has this natural resource that will make it more powerful.

Helen Toner: Exactly. And, again, this is -- each step of the argument is not completely false. Certainly data is important for many types of machine learning systems, though not all. And certainly China does have a larger population and it does seem to have weaker privacy controls in some ways, though not in others.

[...] [29:25] It seems like plenty of other types of data where the US has a huge advantage: Anything to do with military data, whether it be satellite imagery, or data from other sensors that the US government has, the US is just really going to have a big advantage. The whole internet is in English. From what I've read, self-driving car input data tends to be much stronger in the US than in China.

There are just many, many types of relevant data. And what's relevant for any given machine learning system will be different from any other, depending on application. So just to go from consumer data to all data seems like it misses a lot. Aside from the

whole question of "How do the privacy controls actually work?" and "How well can Chinese companies actually integrate data from different sources?" and so on. [...]

Julia Galef: [...] [32:40] Going back for a moment to the US government and their thinking about AI: It has seemed to me that the US government has not been very agent-y when it comes to anticipating the importance of AI. And by agent-y, I mean like planning ahead, taking proactive steps in advance to pursue your goals. Is that your view as well?

Helen Toner: I think, again, with having moved to DC and getting used to things here and so on, it seems like that's kind of true of the US government on basically all fronts. I'm not sure if you disagree.

It's a gigantic bureaucracy that was designed 250 years ago -- that's not completely true, but the blueprints were created 250 years ago -- it's just enormous, and has a huge number of rules and regulations and different people and different agencies and other bodies with different incentives and different plans and different goals. So, I think, to me, it's more like it's kind of a miracle when the US government can be agent-y about something than when it can't. I feel like it's kind of not fair to expect anything else of a structure that is the way that it is. [...]

Julia Galef: [...] [37:30] One thing that people often cite is that China publishes more papers on deep learning than the US does. [...]

Helen Toner: [...] [38:20] How are we counting Chinese versus non-Chinese papers? Because often, it seems to be just doing it via, "Is their last name Chinese?" Which seems like it really is going to miscount. [...]

[...] [39:00] These counts of papers have a hard time saying anything about the quality of the papers involved. You can look at citations, but that's not a perfect metric. But it's better, for sure.

And then, lastly, they rarely say anything about the different incentives that Chinese and non-Chinese academics face in publishing. So, actually, my partner is a chemistry PhD student, and he's currently spending a month in Shanghai. And he mentioned to me spontaneously that it's clear to him, or maybe it got mentioned explicitly, that his professor's salary is dependent on how many papers come out of his lab. So that's just a super different setup. Obviously, in the US, we have plenty of maybe exaggerated incentives for academics to publish papers, but I feel like that's sort of another level.

Best reasons for pessimism about impact of impact measures?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Habryka recently wrote (emphasis mine):

> My inside views on AI Alignment make me think that work on impact measures is *very unlikely* to result in much concrete progress on what I perceive to be core AI Alignment problems, *and I have talked to a variety of other researchers in the field who share that assessment*. I think it's important that this grant not be viewed as an endorsement of the concrete research direction that Alex is pursuing, but only as an endorsement of the higher-level process that he has been using while doing that research.
>
> As such, I think it was a necessary component of this grant that I have talked to other people in AI Alignment whose judgment I trust, who do seem excited about Alex's work on impact measures. I think I would not have recommended this grant, or at least this large of a grant amount, without their endorsement. I think in that case I would have been worried about a risk of diverting attention from what I think are more promising approaches to AI Alignment, and a potential dilution of the field by introducing a set of (to me) somewhat dubious philosophical assumptions.

I'm interested in learning about the intuitions, experience, and facts which inform this pessimism. As such, I'm not interested in making any arguments to the contrary in this post; any pushback I provide in the comments will be with clarification in mind.

There are two reasons you could believe that "work on impact measures is very unlikely to result in much concrete progress on… core AI Alignment problems". First, you might think that the impact measurement problem is intractable, so work is unlikely to make progress. Second, you might think that even a full solution wouldn't be very useful.

Over the course of 5 minutes by the clock, here are the reasons I generated for pessimism (which I either presently agree with or at least find it reasonable that an intelligent critic would raise the concern on the basis of currently-public reasoning):

- Declarative knowledge of a solution to impact measurement probably wouldn't help us do value alignment, figure out embedded agency, etc.
- We want to figure out how to transition to a high-value stable future, and it just isn't clear how impact measures help with that.
- Competitive and social pressures incentivize people to cut corners on safety measures, especially those which add overhead.
    - Computational overhead.
    - Implementation time.
    - Training time, assuming they start with low aggressiveness and dial it up slowly.
- Depending on how "clean" of an impact measure you think we can get, maybe it's way harder to get low-impact agents to do useful things.

    - Maybe we can get a clean one, but only for powerful agents.
    - Maybe the impact measure misses impactful actions if you can't predict at near human level.
- In a world where we know how to build powerful AI but not how to align it (which is actually probably the scenario in which impact measures do the most work), we play a very unfavorable game while we use low-impact agents to somehow transition to a stable, good future: the first person to set the aggressiveness too high, or to discard the impact measure entirely, ends the game.
- In a More realistic tales of doom-esque scenario, it isn't clear how impact helps prevent "gradually drifting off the rails".[1]

---

[1] Paul raised concerns along these lines:

> We'd like to build AI systems that help us resolve the tricky situation that we're in. That help design and enforce agreements to avoid technological risks, build better-aligned AI, negotiate with other actors, predict and manage the impacts of AI, improve our institutions and policy, etc.
>
> I think the default "terrible" scenario is one where increasingly powerful AI makes the world change faster and faster, and makes our situation more and more complex, with humans having less and less of a handle on what is going on or how to steer it in a positive direction. Where we must rely on AI to get anywhere at all, and thereby give up the ability to choose where we are going.
>
> That may ultimately culminate with a catastrophic bang, but if it does it's not going to be because we wanted the AI to have a small impact and it had a large impact. It's probably going to be because we have a very limited idea what is going on, but we don't feel like we have the breathing room to step back and chill out (at least not for long) because we don't believe that everyone else is going to give us time.
>
> If I'm trying to build an AI to help us navigate an increasingly complex and rapidly-changing world, what does "low impact" mean? In what sense do the terrible situations involve higher objective impact than the intended behaviors?
>
> (And realistically I doubt we'll fail at alignment with a bang---it's more likely that the world will just drift off the rails over the course of a few months or years. The intuition that we wouldn't let things go off the rails gradually seems like the same kind of wishful thinking that predicts war or slow-rolling environmental disasters should never happen.)
>
> It seems like "low objective impact" is what we need once we are in the unstable situation where we have the technology to build an AI that would quickly and radically transform the world, but we have all decided not to and so are primarily concerned about radically transforming the world by accident. I think that's a coherent situation to think about and plan for, but we shouldn't mistake it for the mainline. (I personally think it is quite unlikely, and it would definitely be unprecedented, though you could still think it's the best hope if you were very pessimistic about what I consider "mainline" alignment.)

Book review: The Sleepwalkers by Arthur Koestler

> The progress of Science is generally regarded as a kind of clean, rational advance along a straight ascending line; in fact it has followed a zigzag course, at times almost more bewildering than the evolution of political thought. The history of cosmic theories, in particular, may without exaggeration be called a history of collective obsessions and controlled schizophrenias; and the manner in which some of the most fundamental discoveries were arrived at reminds one more of a sleepwalker's performance than an electronic brain's.
>
> Arthur Koestler

The Sleepwalkers is an enlightening history of astronomy from the Ancient Greeks to Newton. It particularly focuses on three characters who shifted scientific consensus from the Ptolemaic geocentric model of the solar system to a heliocentric one: Copernicus, Kepler, and Galileo. And *characters* is the right word, because Koestler digs into their personal quirks and foibles with gusto. If he is to be believed, these three key scientists were all temperamental to the point of self-destructiveness.

Copernicus and Kepler

Copernicus, according to Koestler, was "a man of the Middle Ages: haunted by its anxieties, ridden with its complexes, a timid, conservative cleric, who started the revolution against his will." The origin of his heliocentric system was in fact his attempt to fix an aesthetic flaw which made Ptolemy's system "neither sufficiently absolute nor sufficiently pleasing": that it modelled planets as moving at variable speeds. In doing so, he was inspired by Ancient Greek writings, and in particular references to the heliocentric system that Heraclides of Pontos and Aristarchus of Samos proposed in the 3rd century BC. Copernicus was terrified by the prospect of criticism, so that it took decades for him to publish his conclusions (*On the revolutions of the celestial spheres* only came out just before his death in 1543). But in fact, the publication itself was fairly irrelevant: it was tedious and actually more complicated than existing geocentric systems. As previous astronomers had done, Copernicus explained complex-looking orbits using models in which planets rotated on circles which themselves rotated in circles (known as epicycles). Epicycles were a very powerful modelling tool (in fact, equivalent to Fourier series!) but made theories much uglier and more complex - and Copernicus needed more of them than previous geocentric models. But the important thing was that heliocentrism, an idea which been floating around without much fanfare since the Greeks, was brought to greater prominence - inspiring others, particularly Kepler and Galileo.

Kepler was prone to bursts of passion, alternating between fury and self-deprecation. He was also driven primarily by a philosophical taste verging on mysticism: his first system of astronomy matched the six known planets with the five perfect Platonic solids for aesthetic reasons, and was complete bunk. So too was his theory that the positions of the planets were determined by the ratios of musical harmonies. Yet these instincts kept him pushing, mostly by accident, in the right direction. His breakthrough

came when given access to the meticulous observations of Tycho Brahe (also quite a character: after his nose was chopped off in a duel, he wore a silver replacement). The two had a stormy relationship, but after Tycho's death Kepler slogged through years of calculations to try and pin down the shape of planetary orbits. Through determination and luck (he made several errors which fortunately cancelled out), he eventually formulated his three laws: that planets move in ellipses, with the sun at one focal point; that they travel at variable speeds such that the line between them and the sun sweeps out the same area per unit time; and that the squares of the periods of the planets' revolutions are proportional to the cubes of their mean distances from the sun.

Why the delay?

It's worth highlighting two contributing factors to this very impressive achievement. Firstly, Tycho was *very* well-funded, and took data collection unprecedentedly seriously (note that this was all done before telescopes!). Secondly, Kepler deserves great credit for his persistence in testing his theories against that data, and throwing them out when they failed to meet it. But there's a much broader question of why it took almost 2000 years from the first heliocentric system to the discovery of these laws, and even longer until acceptance of heliocentrism became widespread - why, in Whitehead's words, "in the year 1500 Europe knew less than Archimedes who died in the year 212 BC."

In short, the answer seems to be Plato and Aristotle: as Koestler portrays them, "two frightened men standing in [Plato's] Cave, facing the wall, chained to their places in a catastrophic age, turning their back on the flame of Greece's heroic era, and throwing grotesque shadows which are to haunt mankind for a thousand years and more." This is unduly harsh - and yet how damaging the blind adherence to their ideas which lasted, with few exceptions, until the 16th century! Two of Plato's most influential tenets: that true knowledge cannot be obtained by the study of nature, but must be gained by consideration of the perfect world of Forms and Ideas; and that, for metaphysical reasons, all celestial motion must be in perfect circles at uniform speed. Aristotle then introduced the distinction between the imperfect and changeable earth, and the eternal and unchanging cosmos; also, that each movement was due to an object's *telos*, or purpose. Lastly, God was placed on the outermost sphere, furthest away from the Earth. Their combined influence had astronomers thinking in circles for almost 2000 years! And even the scientists whose work led to the end of these dogmas were still somewhat trapped by them: Kepler spent years trying to make Tycho's data consistent with circular orbits. Only when he felt that he'd conclusively ruled those out did he move on to ovals and ellipses.

More specifically, Koestler identifies five main obstacles to scientific progress between the Greeks and the scientific revolution. "The first was the splitting up of the world into two spheres, and the mental split which resulted from it. The second was the geocentric dogma, the blind eye turned on the promising line of thought which had started with the Pythagoreans and stopped abruptly with Aristarchus of Samos. The third was the dogma of uniform motion in perfect circles. The fourth was the divorcement of science from mathematics. The fifth was the inability to realise that while a body at rest tended to remain at rest, a body in motion tended to remain in motion. The main achievement of the first part of the scientific revolution was the removal of these five cardinal obstacles."

Galileo and Newton

An early step forward was Tycho Brahe's observation of a supernova in 1572, and his demonstration that it wasn't just a comet, but a new (and changing!) phenomenon beyond the orbit of the moon. Then, soon after Kepler published his first two laws in 1609, another blow against the Aristotelean conception of the universe was struck by Galileo's observations of several moons of Jupiter, as well as terrain on the moon and phases of Venus. These were all evidence that the Earth was less special than had been thought. While Galileo hadn't invented the telescope, or been the first to notice the new moons using one, he was the first to bring them to prominence in his book *Sidereal message*. The book faced heavy criticism from other scientists, but eventually became accepted after Kepler threw his support behind it.

This makes Galileo's biggest blunder even more confusing: he simply refused to accept Kepler's system of elliptical orbits, and instead propounded a version of Copernicus' theory which neither was elegant nor matched the data very well. He also dismissed Kepler's correct theory of moon-caused tides in favour of the clearly flawed hypothesis that they were driven by the Earth's motion. This would have been unimportant, if not for his insistence that the Church start reinterpreting scripture based on the evident truth of Copernican heliocentrism, as shown by the existence of tides.* Koestler attributes Galileo's persecution by the Church not to an inevitable clash between science and religion, but rather to Galileo's own bullheadedness in attempting to force a theological surrender without having much evidence, and his tendency to insult and alienate important people. (A particularly damaging example: in his *Dialogue concerning the two chief world systems*, the views of Pope Urban are put into the mouth of a dimwit called Simplicio.) Koestler argues that the Church had a lot of respect for scientists, and in general was willing to reinterpret doctrine based on compelling scientific arguments, but that Galileo's actions made opposition to heliocentrism last much longer than it otherwise would have.**

Nevertheless, we have Galileo to thank for informal precursors to Newton's laws of motion: that objects remain at the same velocity unless acted upon by a force; and that (ignoring air resistance) objects fall with the same constant acceleration regardless of their weight. At this point, science was in a pivotal position. Kepler's laws applied to celestial objects; Galileo's applied to terrestrial objects. There was a growing acceptance that the two were fundamentally the same, but no agreement on what force drove celestial motion, or what planets would do without the influence of the sun, or even what weight meant in a celestial context. People believed Kepler's laws, but didn't understand them: why ellipses? Why an equal area? Ideas were also floating around to do with reciprocal attraction between objects in a way that varied by distance, but nothing concrete - and the introduction of magnetism made people even more confused about which force might be doing what.

Then, of course, came Newton's grand synthesis of gravity in 1687. Consider his thought experiment of firing a cannonball horizontally from a mountain. If fired slowly, it will fall to the earth quite soon. If fired faster, it will curve around the earth a little before falling. But if fired fast enough, it will continually "fall" towards the earth but never hit it: it will be in orbit. Newton was then able to use calculus to derive the fundamental reasons why Kepler's laws must be true (for an intuitive demonstration of why elliptical orbits make sense, try throwing a marble into a cone). Key to this was Newton's acceptance of gravity as "action at a distance", which had always been very controversial. Apart from this, there were almost no cases of two objects affecting each other except by some medium physically between them. The one known

exception, as mentioned, was magnetism - a helpful precedent for building acceptance of Newtonian gravity. Koestler on Newton's importance: "If one had to sum up the history of scientific ideas about the universe in a single sentence, one could only say that up to the seventeenth century our vision was Aristotelian, after that Newtonian. Copernicus and Tycho, Kepler and Galileo, Gilbert and Descartes lived in the no-man's-land between the two."

Science as sleepwalking

As mentioned in the opening quotation, the book is called *The Sleepwalkers* because that whole transition was so confused and messy. Out of Copernicus, Kepler, and Galileo, the first and last never let go of the Platonic ideal of circular orbits. And while Kepler's laws were correct, his actual arguments were riddled with mistakes and contradictions. In addition to the mysticism mentioned above, one anecdote is particularly telling. Kepler believed that elliptical orbits had to result from the combination of two different forces: one linking a planet to the sun, and the other simply acting on the planet itself. This is basically correct: the first is gravity, which pulls planets towards the sun; the second is centrifugal force or inertia, which prevents them from falling inwards. However, in Kepler's theory of physics these are exactly the other way around! He envisaged the sun as providing a force which swept the planets around their rotations, and a planet's intrinsic magnetism as what pulled it towards or away from the sun. His "proof" of his 2nd law was another case of coming to the right conclusion for the wrong reasons. Even Newton, in the midst of his great triumph, had to appeal to God to explain why gravity didn't cause the whole universe to collapse inwards.

Overall I think reading this book is helpful in understanding what it looks like to be very smart and also very confused. It's also notable just how much the leading scientists involved were held back from publishing by concerns about their reputation amongst *other scientists* - and if the ones who we remember now were so worried about that, there were probably many more who succumbed to those worries and who we've therefore never heard of. Lastly, as Koestler highlights several times, the empirical arguments in favour of geocentrism - like "why don't we feel the Earth moving?", "why don't we observe Venus getting closer and further away?", and "why don't we see the stars shifting as we orbit?" - were actually fairly sophisticated and reasonable.*** And given the messiness of Copernicus' system, there simply wasn't enough evidence to conclusively decide in favour of heliocentrism at least until Kepler's ellipses - which were only discovered because Kepler had already devoted his life to the hypothesis. This supports the Kuhnian view of scientific revolutions as driven to a significant extent by personal taste in paradigms. *The Sleepwalkers* was actually published just three years before *The Structure of Scientific Revolutions*, and in the epilogue Koestler discusses some very similar ideas:

"The philosophy of nature evolved by occasional leaps and bounds alternating with delusional pursuits, *culs-de-sac,* regressions, periods of blindness, and amnesia. The great discoveries which determined its course were sometimes the unexpected by-products of a chase after quite different hares. … There occur in biological evolution periods of crisis and transition when there is a rapid, almost explosive branching out in all directions, often resulting in a radical change in the dominant trend of development. The same kind of thing seems to have happened in the evolution of thought at critical periods like the 6th century BC or the 17th AD. ... 'Intellectual progress' has, as it were, linear associations - a continuous curve, a steadily rising

water level; whereas ‘evolution’ is known to be a wasteful, fumbling process characterised by sudden mutations of unknown cause, by the slow grinding of selection, and by the dead-ends of over-specialisation and rigid inadaptability."

And finally, a prescient passage on the increasing power of science:

"Thus within the foreseeable future, man will either destroy himself or take off from the stars. It is doubtful whether reasoned argument will play any significant part in the ultimate decision, but if it does, a clearer insight into the evolution of ideas which led to the present predicament may be of some value. The muddle of inspiration and delusion, of visionary insight and dogmatic blindness, of millennial obsessions and disciplined double-think, which this narrative has tried to retrace, may serve as a cautionary tale against the *hubris* of science - or rather of the philosophical outlook based on it. The dials on our laboratory panels are turning into another version of the shadows in the cave."

* One important concept here is the difference between a model which matches observed data (thereby “saving the phenomenon”) and a hypothesis which is claimed to be literally true. My impression is that geocentric models with epicycles were generally considered part of the former category, because while they matched the data pretty well, there was no good explanation of why planets would move *as if* they were on wheels attached to wheels, when the existence of such wheels in the sky would be absurd. The Church also had no problem with heliocentrism when presented as the former (which Galileo refused to do). Note that the question of when we should treat theories as descriptive models or literal truth is still a key question in philosophy of science (now under the headings of *scientific anti-realism* versus *scientific realism).*

** Note that while Koestler’s scholarship generally seems meticulous, with copious quotes from original sources, I’m still slightly skeptical about this argument - particularly after reading the epilogue, which displays a pro-religion bias and credulity towards ESP and time travel. See also this argument that even easily-avoidable censorship is dangerous.

*** For another excellent perspective on this, see Jacob Lagerros' essay here.

Any rebuttals of Christiano and AI Impacts on takeoff speeds?

14 months ago, Paul Christiano and AI Impacts both published forceful and well-received take-downs of many arguments for fast (discontinuous) takeoff. I haven't seen any rebuttals that are written by established researchers, longer than comments, or otherwise convincing. The longer there is no response, the less weight I put on the outside view that proponents of fast takeoff may be right.

Where are the rebuttals? Did I miss them? Is the debate decided? Did nobody have time or motivation to write something? Is the topic too hard to explain?

Why rebuttals would be useful:

-Give the community a sense of the extent of expert disagreement to form outside views.

-Prioritization in AI policy, and to a lesser extent safety, depends on the likelihood of discontinuous progress. We may have more leverage in such cases, but this could be overwhelmed if the probability is low.

-Motivate more people to work on MIRI's research which seems more important to solve early if there is fast takeoff.

Where to Draw the Boundaries?

Followup to: Where to Draw the Boundary?

Figuring where to cut reality in order to carve along the joints—figuring which things are similar to each other, which things are clustered together: this *is the problem worthy of a rationalist. It is what people* should *be trying to do, when they set out in search of the floating essence of a word.*

Once upon a time it was thought that the word "fish" included dolphins ...

The one comes to you and says:

> The list: `{salmon, guppies, sharks, dolphins, trout}` is just a list—you can't say that a list is *wrong*. You draw category boundaries in specific ways to capture tradeoffs you care about: sailors in the ancient world wanted a word to describe the swimming finned creatures that they saw in the sea, which included salmon, guppies, sharks—and dolphins. That grouping may not be the one favored by modern evolutionary biologists, but an alternative categorization system is not an error, and borders are not objectively true or false. You're not standing in defense of truth if you insist on a word, brought explicitly into question, being used with some particular meaning. So my definition of *fish* cannot possibly be 'wrong,' as you claim. I can define a word any way I want—in accordance with my values!

So, there is a legitimate complaint here. It's true that sailors in the ancient world had a legitimate reason to want a word in their language whose extension was `{salmon, guppies, sharks, dolphins, ...}`. (And modern scholars writing a translation for present-day English speakers might even translate that word as *fish*, because most members of that category are what we would call fish.) It indeed would not necessarily be helping the sailors to tell them that they need to exclude dolphins from the extension of that word, and instead include dolphins in the extension of their word for `{monkeys, squirrels, horses ...}`. Likewise, most modern biologists have little use for a word that groups dolphins and guppies together.

When rationalists say that definitions can be wrong, we don't mean that there's a *unique* category boundary that is the True floating essence of a word, and that all other possible boundaries are wrong. We mean that in order for a proposed category boundary to *not* be wrong, it needs to capture some statistical structure in reality, even if reality is surprisingly detailed and there can be *more than one* such structure.

The reason that the sailor's concept of *water-dwelling animals* isn't necessarily wrong (at least within a particular domain of application) is because dolphins and fish actually *do* have things in common due to convergent evolution, despite their differing ancestries. If we've been told that "dolphins" are water-dwellers, we can *correctly* predict that they're likely to have fins and a hydrodynamic shape, even if we've never seen a dolphin ourselves. On the other hand, if we predict that dolphins probably lay eggs because 97% of known fish species are oviparous, we'd get the *wrong answer*.

A standard technique for understanding why some objects belong in the same "category" is to (pretend that we can) visualize objects as existing in a very-high-dimensional configuration space, but this "Thingspace" isn't particularly well-defined: we want to map every property of an object to a dimension in our abstract space, but it's not clear how one would enumerate all possible "properties." But this isn't a major

concern: we can form a space with *whatever* properties or variables we happen to be interested in. Different choices of properties correspond to different cross sections of the grander Thingspace. Excluding properties from a collection would result in a "thinner", lower-dimensional subspace of the space defined by the original collection of properties, which would in turn be a subspace of grander Thingspace, just as a line is a subspace of a plane, and a plane is a subspace of three-dimensional space.

Concerning dolphins: there would be a cluster of water-dwelling animals in the subspace of dimensions that water-dwelling animals are similar on, and a cluster of mammals in the subspace of dimensions that mammals are similar on, and dolphins would belong to *both* of them, just as the vector [1.1, 2.1, 9.1, 10.2] in the four-dimensional vector space $\mathbb{R}^4$ is simultaneously close to [1, 2, 2, 1] in the subspace spanned by x_1 and x_2, *and* close to [8, 9, 9, 10] in the subspace spanned by x_3 and x_4.

Humans are already functioning intelligences (well, sort of), so the categories that humans propose of their own accord won't be *maximally* wrong: no one would try to propose a word for "configurations of matter that match any of these 29,122 five-megabyte descriptions but have no other particular properties in common." (Indeed, because we are not-superexponentially-vast minds that evolved to function in a simple, ordered universe, it actually takes some ingenuity to construct a category *that* wrong.)

This leaves aspiring instructors of rationality in something of a predicament: in order to *teach* people how categories can be more or (ahem) less wrong, you need some sort of illustrative example, but since the most natural illustrative examples won't be *maximally* wrong, some people might fail to appreciate the lesson, leaving one of your students to fill in the gap in your lecture series eleven years later.

The *pedagogical* function of telling people to "stop playing nitwit games and admit that dolphins don't belong on the fish list" is to point out that, without *denying* the obvious similarities that motivated the initial categorization `{salmon, guppies, sharks, dolphins, trout, ...}`, there is *more structure* in the world: to maximize the (logarithm of the) probability your world-model assigns to your observations of dolphins, you need to take into consideration the many aspects of reality in which the grouping `{monkeys, squirrels, dolphins, horses ...}` makes more sense. To the extent that relying on the initial category guess would result in a worse Bayes-score, we might say that that category is "wrong." It might have been "good enough" for the purposes of the sailors of yore, but as humanity has learned more, as our model of Thingspace has expanded with more dimensions and more details, we can see the ways in which the original map failed to carve reality at the joints.

The one replies:

> But reality doesn't come with its joints pre-labeled. Questions about how to draw category boundaries are best understood as questions about values or priorities rather than about the actual content of the actual world. I can call dolphins "fish" and go on to make just as accurate predictions about dolphins as you can. Everything we identify as a joint is only a joint because we care about it.

No. Everything we identify as a joint is a joint not "because we care about it", but because it *helps us think about* the things we care about.

Which dimensions of Thingspace you bother paying attention to might depend on your values, and the clusters returned by your brain's similarity-detection algorithms might "split" or "collapse" according to which subspace you're looking at. But in order for your map to be *useful* in the service of your values, it needs to reflect the statistical structure of things in the territory—which depends on the territory, not your values.

There is an *important difference* between "not including mountains on a map because it's a political map that doesn't show any mountains" and "not including Mt. Everest on a geographic map, because my sister died trying to climb Everest and seeing it on the map would make me feel sad."

There is an *important difference* between "identifying this pill as not being 'poison' allows me to focus my uncertainty about what I'll observe after administering the pill to a human (even if most possible minds have never seen a 'human' and would never waste cycles imagining administering the pill to one)" and "identifying this pill as not being 'poison', because if I publicly called it 'poison', then the manufacturer of the pill might sue me."

There is an *important difference* between having a utility function defined over a statistical model's *performance* against specific real-world data (even if another mind with different values would be interested in different data), and having a utility function defined over features of *the model itself*.

Remember how appealing to the dictionary is irrational when the *actual* motivation for an argument is about whether to infer a property on the basis of category-membership? But at *least* the dictionary has the virtue of documenting typical usage of our shared communication signals: you can at least see how "You're defecting from common usage" might *feel* like a sensible thing to say, even if one's true rejection lies elsewhere. In contrast, this motion of appealing to *personal values* (!?!) is *so* deranged that Yudkowsky apparently didn't even realize in 2008 that he might need to warn us against it!

You *can't* change the categories your mind *actually* uses and still perform as well on prediction tasks—although you can change your *verbally reported* categories, much as how one can verbally report "believing" in an invisible, inaudible, flour-permeable dragon in one's garage without having any false anticipations-of-experience about the garage.

This may be easier to see with a simple *numerical* example.

Suppose we have some entities that exist in the three-dimensional vector space $\mathbb{R}^3$. There's one cluster of entities centered at [1, 2, 3], and we call those entities Foos, and there's another cluster of entities centered at [2, 4, 6], which we call Quuxes.

The one comes and says, "Well, I'm going redefine the meaning of 'Foo' such that it also includes the things near [2, 4, 6] as well as the Foos-with-respect-to-the-old-definition, and you can't say my new definition is wrong, because if I observe [2, _, _] (where the underscores represent yet-unobserved variables), I'm going to categorize that entity as a Foo but still predict that the unobserved variables are 4 and 6, *so there*."

But if the one were *actually using* the new concept of Foo *internally* and not just *saying the words* "categorize it as a Foo", they *wouldn't* predict 4 and 6! They'd predict 3 and 4.5, because those are the average values of a generic Foo-with-respect-to-the-new-definition in the 2nd and 3rd coordinates (because (2+4)/2 = 6/2

= 3 and (3+6)/2 = 9/2 = 4.5). (The already-observed 2 in the first coordinate isn't average, but by conditional independence, that only affects our prediction of the other two variables *by means* of its effect on our "prediction" of category-membership.) The cluster-structure knowledge that "entities for which $x_1 \approx 2$, also tend to have $x_2 \approx 4$ and $x_3 \approx 6$" needs to be represented *somewhere* in the one's mind *in order to get the right answer*. And given that that knowledge needs to be represented, it might also be useful to have a *word* for "the things near [2, 4, 6]" in order to efficiently share that knowledge with others.

Of course, there isn't going to be a *unique* way to encode the knowledge into natural language: there's no reason the word/symbol "Foo" needs to represent "the stuff near [1, 2, 3]" rather than "both the stuff near [1, 2, 3] and also the stuff near [2, 4, 6]". And you might very well indeed want a short word like "Foo" that encompasses both clusters, for example, if you want to contrast them to another cluster much farther away, or if you're mostly interested in x_1 and the difference between $x_1 \approx 1$ and $x_1 \approx 2$ doesn't seem large enough to notice.

But if speakers of particular language were *already* using "Foo" to specifically talk about the stuff near [1, 2, 3], then you can't swap in a new definition of "Foo" without *changing the truth values* of sentences involving the word "Foo." Or rather: sentences involving Foo-with-respect-to-the-old-definition are *different* propositions from sentences involving Foo-with-respect-to-the-new-definition, even if they get written down using the same symbols in the same order.

Naturally, all this becomes much more complicated as we move away from the simplest idealized examples.

For example, if the points are more evenly distributed in configuration space rather than belonging to cleanly-distinguishable clusters, then essentialist "X is a Y" cognitive algorithms perform less well, and we get Sorites paradox-like situations, where we know *roughly* what we mean by a word, but are confronted with real-world (not merely hypothetical) edge cases that we're not sure how to classify.

Or it might not be obvious which dimensions of Thingspace are most relevant.

Or there might be social or psychological forces anchoring word usages on identifiable Schelling points that are easy for different people to *agree* upon, even at the cost of some statistical "fit."

We could go on listing more such complications, where we seem to be faced with somewhat arbitrary choices about how to describe the world in language. But the fundamental thing is this: *the map is not the territory*. Arbitrariness in the map (what color should Texas be?) doesn't correspond to arbitrariness in the territory. Where the structure of human natural language doesn't fit the structure in reality—where we're not sure whether to say that a sufficiently small collection of sand "is a heap", because we don't know how to specify the positions of the individual grains of sand, or compute that the collection has a Standard Heap-ness Coefficient of 0.64—that's just a *bug* in our human power of vibratory telepathy. You can exploit the bug to confuse humans, but that doesn't change reality.

Sometimes we might *wish* that something to belonged to a category that it doesn't (with respect to the category boundaries that we would ordinarily use), so it's tempting to avert our attention from this painful reality with appeal-to-arbitrariness language-lawyering, selectively applying our philosophy-of-language skills to pretend that we can define a word any way we want with no consequences. ("I'm not late!—

well, okay, we agree that I arrived half an hour after the scheduled start time, but whether I was *late* depends on how you choose to draw the category boundaries of 'late', which is subjective.")

For this reason it is said that knowing about philosophy of language can hurt people. Those who know that words don't have intrinsic definitions, but *don't* know (or have seemingly forgotten) about the three or six dozen optimality criteria governing the use of words, can easily fashion themselves a Fully General Counterargument against *any* claim of the form "X is a Y"—

> Y doesn't unambiguously refer to the thing you're trying to point at. There's no Platonic essence of Y-ness: once we know any particular fact about X we want to know, there's no question left to ask. Clearly, you don't understand how words work, therefore I don't need to consider whether there are any non-ontologically-confused reasons for someone to say "X is a Y."

Isolated demands for rigor are great for winning arguments against humans who aren't as philosophically sophisticated as you, but the evolved systems of perception and language by which humans process and communicate information about reality, *predate* the Sequences. Every claim that X is a Y is an expression of *cognitive work* that cannot simply be dismissed just because most claimants doesn't know *how* they work. Platonic essences are just the limiting case as the overlap between clusters in Thingspace goes to zero.

You should *never* say, "The choice of word is arbitrary; therefore I can say whatever I want"—which amounts to, "The choice of category is arbitrary, therefore I can believe whatever I want." If the choice were *really* arbitrary, you would be satisfied with the choice being *made* arbitrarily: by flipping a coin, or calling a random number generator. (It doesn't matter which.) Whatever criterion your brain is using to decide which word or belief you want, *is* your non-arbitrary reason.

If what you want isn't currently true in reality, maybe there's some action you could take to make it *become* true. To search for that action, you're going to need accurate beliefs about what reality is *currently* like. To enlist the help of others in your planning, you're going to need precise terminology to *communicate* accurate beliefs about what reality is currently like. Even when—*especially* when—the current reality is inconvenient.

Even when it hurts.

(Oh, and if you're actually trying to optimize other people's models of the world, rather than the world itself—you could just *lie*, rather than playing clever category-gerrymandering mind games. It would be a lot simpler!)

Imagine that you've had a peculiar job in a peculiar factory for a long time. After many mind-numbing years of sorting bleggs and rubes all day and enduring being trolled by Susan the Senior Sorter and her evil sense of humor, you finally work up the courage to ask Bob the Big Boss for a promotion.

"Sure," Bob says. "Starting tomorrow, you're our new Vice President of Sorting!"

"Wow, this is amazing," you say. "I don't know what to ask first! What will my new responsibilities be?"

"Oh, your responsibilities will be the same: sort bleggs and rubes every Monday through Friday from 9 *a.m.* to 5 *p.m.*"

You frown. "Okay. But Vice Presidents get paid a lot, right? What will my salary be?"

"Still $9.50 hourly wages, just like now."

You grimace. "O–*kay*. But Vice Presidents get more authority, right? Will I be someone's boss?"

"No, you'll still report to Susan, just like now."

You snort. "A Vice President, reporting to a mere Senior Sorter?"

"Oh, no," says Bob. "Susan is *also* getting promoted—to *Senior* Vice President of Sorting!"

You lose it. "Bob, this is *bullshit*. When you said I was getting promoted to Vice President, that created a bunch of probabilistic expectations in my mind: you made me *anticipate* getting new challenges, more money, and more authority, and then you reveal that you're just slapping an inflated title on the same old dead-end job. It's like handing me a blegg, and then saying that it's a rube that just happens to be blue, furry, and egg-shaped ... or telling me you have a dragon in your garage, except that it's an invisible, silent dragon that doesn't breathe. You may *think* you're being kind to me asking me to believe in an unfalsifiable promotion, but when you replace the symbol with the substance, it's actually just cruel. *Stop fucking with my head!* ... sir."

Bob looks offended. "This promotion isn't *unfalsifiable*," he says. "It *says*, 'Vice President of Sorting' right here on the employee roster. That's an sensory experience that you can make falsifiable predictions about. I'll even get you business cards that say, 'Vice President of Sorting.' That's another falsifiable prediction. Using language in a way *you* dislike is not lying. The propositions you claim false—about new job tasks, increased pay and authority—is not what the title is meant to convey, and this is known to everyone involved; it is not a secret."

Bob *kind of* has a point. It's tempting to argue that things like titles and names are part of the map, not the territory. Unless the name is written down. Or spoken aloud (instantiated in sound waves). Or *thought about* (instantiated in neurons). The map is *part* of the territory: insisting that the title isn't part of the "job" and therefore violates the maxim that meaningful beliefs must have testable consequences, doesn't quite work. Observing the title on the employee roster indeed tightly constrains your anticipated experience of the title on the business card. So, that's a non-gerrymandered, predictively useful category ... right? What is there for a rationalist to complain about?

To see the problem, we must turn to information theory.

Let's imagine that an abstract Job has four binary properties that can either be `high` or `low`—task complexity, pay, authority, and prestige of title—forming a four-dimensional Jobspace. Suppose that two-thirds of Jobs have `{complexity: low, pay: low, authority: low, title: low}` (which we'll write more briefly as [low, low, low, low]) and the remaining one-third have `{complexity: high, pay: high, authority: high, title: high}` (which we'll write as [high, high, high, high]).

Task complexity and authority are hard to perceive outside of the company, and pay is only negotiated after an offer is made, so people deciding to seek a Job can only make decisions based the Job's title: but that's fine, because in the scenario described, you can infer any of the other properties from the title with certainty. Because the properties are either *all* low or *all* high, the joint entropy of title and any other property is going to have the same value as either of the individual property entropies, namely $\frac{2}{3} \log_2 3/2 + \frac{1}{3} \log_2 3 \approx 0.918$ bits.

But since H(pay) = H(title) = H(pay, title), then the mutual information I(pay; title) has the same value, because I(pay; title) = H(pay) + H(title) − H(pay, title) by definition.

Then suppose a *lot* of companies get Bob's bright idea: half of the Jobs that used to occupy the point [low, low, low, low] in Jobspace, get their title coordinate changed to high. So now one-third of the Jobs are at [low, low, low, low], another third are at [low, low, low, high], and the remaining third are at [high, high, high, high]. What happens to the mutual information I(pay; title)?

I(pay; title) = H(pay) + H(title) − H(pay, title)
= (⅔ log 3/2 + ⅓ log 3) + (⅔ log 3/2 + ⅓ log 3) − 3(⅓ log 3)
= 4/3 log 3/2 + 2/3 log 3 − log 3 ≈ 0.2516 bits.

It went down! Bob and his analogues, having observed that employees and Job-seekers prefer Jobs with high-prestige titles, *thought* they were being benevolent by making more Jobs have the desired titles. And perhaps they have helped savvy employees who can arbitrage the gap between the new and old worlds by being able to put "Vice President" on their resumés when searching for a new Job.

But from the perspective of people who wanted to use titles as an easily-communicable correlate of the other features of a Job, all that's actually been accomplished is *making language less useful*.

In view of the preceding discussion, to "37 Ways That Words Can Be Wrong", we might wish to append, "38. Your definition draws a boundary around a cluster in an inappropriately 'thin' subspace of Thingspace that excludes relevant variables, resulting in fallacies of compression."

Miyamoto Musashi is quoted:

> The primary thing when you take a sword in your hands is your intention to cut the enemy, whatever the means. Whenever you parry, hit, spring, strike or touch the enemy's cutting sword, you must cut the enemy in the same movement. It is essential to attain this. If you think only of hitting, springing, striking or touching the enemy, you will not be able actually to cut him.

Similarly, the primary thing when you take a word in your lips is your intention to reflect the territory, whatever the means. Whenever you categorize, label, name, define, or draw boundaries, you must cut through to the correct answer in the same movement. If you think only of categorizing, labeling, naming, defining, or drawing boundaries, you will not be able actually to reflect the territory.

Do not ask whether there's a rule of rationality saying that you shouldn't call dolphins fish. Ask whether dolphins are fish.

And if you speak overmuch of the Way you will not attain it.

(Thanks to Alicorn, Sarah Constantin, Ben Hoffman, Zvi Mowshowitz, Jessica Taylor, and Michael Vassar for feedback.)

Robin Hanson on Simple, Evidence Backed Models

This is a linkpost for http://www.overcomingbias.com/2019/04/publish-tax-returns.html

This is link to a Hanson post that is primarily about tax returns, and whether they should be public. It was an interesting foray into the considerations that bear on that topic.

But I was particularly interested in the opening paragraph, which had a useful lens:

> Our simplest model of an economy is: supply and demand. This model has many simple implications for policy. Now we know of many much more complicated economic models, which often have quite different policy implications. But often we are not sure which more complex models actually apply well to any given situation. So we have to worry that people favor more complex models mainly to justify their preferred policies. Knowing this pushes me toward recommending the policies implied by supply and demand, unless I see unusually clear evidence to support a different economic model. (FYI, the evidence that fixed costs exists seems plenty clear, so I really mean supply & demand with fixed costs.)
>
> In our simplest modes of information, people are better off when they have more information, and also when information is distributed more symmetrically. There is a vast world of much more complicated models, but it is often hard to tell which more complex models apply well to given situations, and many probably favor particular models to justify preferred policies. So as with supply and demand, this uncertainty pushes me to favor the simplest info models, and their policy recommendations favoring more info and more symmetric info.

Why does category theory exist?

- Why did someone bother coming up with it?
- Why did they think it might be able to give us useful insights?
- And what interesting insights into reality does it give us?

Counterspells

Back in the early days of the internet, when New Atheism was king, people would collect and catalogue various logical fallacies. Then, when getting into an argument on some forum, they would identify and call out all the logical fallacies being used by their opponent; Ad Hominem! Appeal to Emotion! Tu Quoque! Post hoc, ergo propter hoc!

These days we tend to know better. Appealing to these lists of logical fallacies and their Latin names doesn't help you, and it doesn't help the conversation. The problem is that to the person you're talking to, throwing around the names of these fallacies just sounds like an Argument from Authority—

Wait! I mean, it just sounds like you think you're smarter or better educated than they are, which isn't relevant to the argument. If you really had something to contribute, you would explain the problem with their reasoning, rather than just throwing out a Latin term they've never heard of. Even if they *have* heard of it, you should still take the time to tell them why you think it applies to what they said.

Logical fallacies are all sound complaints. But when disagreeing with someone, they deserve to hear the arguments those complaints are based on. More than that, you're not going to convince someone of your position unless you actually try to convince them.

Fortunately, logical fallacies are *such* bad arguments that most of them can be rebutted in a few simple sentences. And the form of each fallacy tends to be so consistent that these rebuttals can be highly formulaic.

In accordance with the Rationalist tradition that requires everything to have a nerdy sci-fi or fantasy name, I am calling these formulaic rebuttals **Counterspells**.

Like I mentioned, I know that by now everyone knows how to avoid using these fallacies themselves. Most of us can also recognize when other people are using these bad arguments against us. Somehow the next step doesn't come as easily.

You might think that an intelligent person, recognizing a bad argument they've seen a hundred times before, would be able to provide a knock-down counterargument. But for me at least, I still regularly crash and burn in my attempts to extemporize rebuttals. I see a lot of you all make the same problem.

Some writers have proposed some great Counterspells, but no one has collected them. To begin with, I'm going to review some excellent Counterspells by Paul Graham and Scott Alexander. I've lifted many of their examples directly, as these examples were carefully considered and strongly written; check out the links to see the original pieces.

Graham's Counterspells

Paul Graham's How to Disagree is an old favorite. Part of what I find so appealing about this essay is how Graham not only lays out several forms of unproductive

disagreement, but includes clear examples of the sort of thing one should say in response.

Without further ado, Graham's Counterspells:

Ad Hominem

Forms:

- If there's something wrong with [person's] argument, what is it? If there isn't, what difference does it make that [person] is a [member of category]?
- The question is whether [person] is correct or not. If [person's lack of authority or other issue] caused them to make mistakes, what are the mistakes? If there aren't mistakes, [person's lack of authority or other issue] isn't a problem.

Examples:

If there's something wrong with the senator's argument, you should say what it is; and if there isn't, what difference does it make that he's a senator?

The question is whether I'm correct or not. If the fact that I'm just a college student caused me to make mistakes, what mistakes did I make? If there aren't mistakes, then my age isn't a problem.

Response to Tone

Forms:

- It matters much more whether [person] is wrong or right than what their tone is.
- I agree about [issue with tone], but I care much more about if [person] is incorrect somewhere. Could you point out where you think the argument is wrong?

Examples:

It's clear that OP is pissed off about this issue, but it matters more whether OP is wrong or right than what their tone is. Do you think they're wrong, or just an asshole?

Ok, maybe I was too flippant, but I still think my points are correct. If you think I'm actually wrong, could you point out where you disagree?

Contradiction

Forms:

- I think [person] is just [stating the opposing case], without explaining why [person] thinks it's correct. What evidence that convinced [person]?
- It's good to know [other person's position], but I actually still disagree. But I need to understand what evidence makes you think [position].

Examples:

I think you're just stating that you oppose gun control, without explaining why you think it should be opposed. What is it exactly about gun control that makes you so strongly against these policies?

Ok, it's good to know that you're concerned about genetically engineered crops specifically, but I actually still disagree. What evidence makes you think that these crops aren't safe?

Alexander's Counterspells

Noncentral Fallacy

One of my favorite pieces by Alexander is his description of his candidate for the Worst Argument In The World, what he calls the Noncentral Fallacy.

One thing that I like in particular about this piece is that he clearly explains how one should respond to this type of argument. This sets us up nicely for Counterspells.

Forms:

- Yes, so what? [Object of discussion] is [member of a category]. But there are all sorts of other things we know about [object of discussion], and I think they're also relevant to the question at hand. [Give example of other relevant information if possible.]
- Normally when we think about [issue] we mean something like [typical/central example of issue]. [This case] differs in that [ways in which this case is different].
- Obviously [what we're discussing] is an example of [issue]. The typical example of [issue] — [give typical example of issue] — is pretty [bad/good/etc.]. But a lot of the reasons [typical example of issue] — [list of 2-3 ways typical example of issue is bad/good/whatever] — don't apply to the wildly atypical case of [what we're discussing].

Examples:

The typical case of murder is Charles Manson breaking into a house and shooting someone. Abortion differs in that the victim is an embryo or fetus with less biological complexity and intelligence than the average rabbit.

The typical example of theft is someone mugging you in a dark alley and taking your pocketbook. Taxation technically qualifies as theft if you define the latter as "taking someone's money through implied threat of force", but it also differs from dark-alley-mugging in several important ways, like that it's levied by a democratically elected government, that it's supposed to be spent on useful programs, and that it's collected in an orderly and predictable fashion. These differences seem to be important enough that most people support taxation even though they don't support dark-alley-muggings.

Well, obviously*. That's kind of the* point*. And the typical example of racial discrimination - the Ku Klux Klan burning your house down or something - is pretty bad. But a lot of the reasons KKK-house-burning is bad - living in fear, locking downtrodden groups into a cycle of poverty, totally locking qualified people out of any job - don't apply to the wildly atypical case of affirmative action.*

Alexander has an essay called Varieties of Argumentative Experience, written as something of a response to Graham's essay. This is excellent for my purposes, since it means that it has a very similar form. Here are some Counterspells drawn from this source.

Gotchas

Gotchas don't share as much structure as other things on this list, so the form of their Counterspells are not as reliable as some others. I think that Alexander would also tell me that Gotchas aren't worth engaging with, and maybe he's right. But I hold out hope for people who are still playing gotcha, and a calm response can help to refocus the conversation on the issues, so here is an attempt.

You'll notice that my approach here is to deadpan accept the fiction that they are making an honest contribution. Other approaches might also be possible.

Forms:

- Well, [things] can vary along dimensions other than [thing being pointed out].
- Yeah, but I still support [thing] because it has [these other consequences].

Examples:

Well, I don't want to move to Cuba, but that's because governments can vary along dimensions other than how big they are, and even though Cuba has a big government there are a lot of things about it that I don't like.

Yeah, but I still support serious gun control measures because I think there are bigger issues than criminals having guns, like suicide rates and murders by private citizens.

Single Facts

Forms:

- [Fact] admittedly does support your argument, but things with some bad features can be good overall. [Relevant example, if possible.]
- [Fact] is a good point, but the issue is larger than that, and I think [fact] is outweighed by [facts that outweigh it].
- Even if we all agree that [fact] is true, you're only pointing out one [bad/good] quality of [thing]. I don't think it's such a big deal. Why is [fact] really important?

Examples:

Trump lying about his grades admittedly does support your argument that he's sometimes dishonest, but things with some bad features can be good overall. Trump can be dishonest as a businessman sometimes, but he has other good qualities.

Obviously foreign intervention does cost American lives, and that's something I worry about too. But the issue is larger than that, and I think we should be prepared to make these sacrifices sometimes — things like genocide and religious dictatorships are much worse.

Even if we all agree that Hillary is crap at email security, you're only pointing out one bad quality of hers. I don't think it's such a big deal. Why is her email security cred so important?

Single Studies

There are a lot of ways in which sharing a single study can be a nonproductive contribution to a discussion, so here are a few varieties of Counterspell.

Forms:

- On any controversial issue, there are usually many peer-reviewed studies supporting each side. In just [however much time you spent searching] I was able to find [X number] articles with the exact opposite finding. If we want to answer this question, we need to take a look at the literature as a whole.
- There are a lot of ways in which a study can be wrong, and it's not always obvious when they are. Sometimes honest researchers make a mistake, or just get unlucky. Is there any other evidence out there for [thing] besides this study?
- I think this study investigates [a much weaker subproblem; be as specific as possible] rather than [the larger problem].

Examples:

On any controversial issue, there are usually many peer-reviewed studies supporting each side. In just 15 minutes, I was able to find 7 articles (links here) that conclude that raising the minimum wage decreases *unemployment. If we want to answer this question, we need to take a look at the literature as a whole.*

I think this study investigates whether trigger warnings affect certain beliefs about trauma in the very short term rather than whether trigger warnings are helpful or harmful in college courses, which is the thing most people actually care about.

Isolated Demand for Rigor

This actually can be countered in a variety of interesting ways; see this essay for a more complete treatment. Most examples lifted from there.

Forms:

- This is wrong because [clear counterexample to the demand for rigor]; [proposed rule] is a fake rule we never apply to anything else.
- Presumably you think [thing that doesn't pass the demand for rigor]. So why does [issue] have to justify itself according to rigorous criteria that nothing could possibly pass?
- So you think that [rule stated explicitly]? Wouldn't that imply [weird consequence they probably don't support]?
- You seem to be happy to [do thing that breaks the demanded rigor], but you switch to a position that [other thing isn't ok because of demand for rigor]. Why the difference in opinion?

Examples:

Republicans have also been against leaders who presided over good economies and presumably thought this was a reasonable thing to do; “it’s impossible to honestly oppose someone even when there’s a good economy” is a fake rule we never apply to anything else.

Presumably there are lots of government programs you support (maybe PBS?), but you would never dream of demanding that we defund them in the hopes of donating the money to malaria prevention. But since you don't support air strikes, suddenly that plan has to justify itself according to rigorous criteria that no government program that exists could possibly pass.

So you think that no one should ever be forced to pay for something they don’t like? Wouldn't that imply liberals shouldn’t have to pay for wars? That seems like a weird consequence you probably don't support.

You're probably happy to talk about speed and blood pressure and comas and the crime rate, but somehow you switch to a position that we can’t talk about IQ at all unless we have a perfect factor-analytical proof of its obeying certain statistical rules. These seem equally rigorous, so why are you ok with one and not the other?

Disputing Definitions

Alexander correctly notes that this is one of the hardest issues to rebut; both because it's easy to be seduced by this sort of argument, and because thinking critically about words and definitions is a very difficult skill to learn, a point not readily communicated in a few sentences.

Even so, here are my own attempts at a Counterspell for this issue:

Forms:

- I'm not sure it matters whether or not [thing] is a member of [category]. People might even disagree with the definition of [category]. [Give example of definition drift if possible.] I think we both care a lot more about things like [actual issues with material consequences].
- You can define [category] as [your definition] if you want, but that's not usually what people mean by [category]. I don't think factual or moral questions depend on how we use words.
- You can define [category] as [your definition] if you want, but I'm much more concerned with [practical issue unrelated to the definition of a word].

(Note that this has a lot of overlap with the Noncentral Fallacy.)

Examples:

I'm not sure it matters whether or not being transgender is a mental illness. People might even disagree how to define what is and isn't a mental illness. Some people still don't think depression would count! I think we both care a lot more about things like whether or not transgender people should be allowed to change their legal gender, and whether they face discrimination.

You can define communism so that the USSR doesn't count if you want, but that's not usually what people mean by communism. I think the factual and moral issues stand, regardless of what words you use.

You can define war as "a formal state of armed conflict between state governments" if you want, but I'm much more concerned with the fact that people are dying in this conflict, whatever we call it.

Why can't the chef call a tomato a vegetable? He just wants to think about whether or not to put it in a salad or on a burger. What's wrong with that?

Other Counterspells

And here are a few of my own design, for some common fallacies:

Argument from Authority

Forms:

- Just because [source] is [better educated/more famous/more expert/smarter] than me, isn't relevant to [the issue]. Sometimes experts are wrong and make mistakes. What arguments does [source] make that convince you of that position?

Examples:

Pinker being a Harvard professor and generally a smart guy isn't relevant to his position on AI threat. Sometimes experts are wrong and make mistakes. What part of Pinker's arguments do you think refute my position?

Straw Man

Forms:

- Well I can't speak for everyone else, but [straw man] isn't what I believe about this case. Are you sure you understand what my position is? [If charitable, attempt to clarify your position.]
- ***(If you are very confident)*** I don't think anyone really believes [straw man]. Personally I believe [state your position]. Do you disagree with that?

Examples:

Well I can't speak for everyone else, but I don't actually want to ban all private ownership of guns. Are you sure you understand what my position is? All I'm suggesting is that we enforce the laws we have, and close some loopholes related to purchasing firearms at gun shows.

Tu Quoque

Forms:

- It's possible be against something and still sometimes participate in it. I may be a hypocrite about [whatever], but that doesn't make me wrong.

Examples:

It's possible be against something and still sometimes participate in it. Sure, it's hypocritical for me to eat meat, but that doesn't make me wrong about animal rights.

Sure, it's hypocritical for me to smoke and tell you not to, but I'm still right that it would be bad for your health, and I can still want you not to start!

Appeal to Popularity

Forms:

- The majority has been wrong all the time. [Pick one of the many examples, one that your audience will agree with.] I don't care who [agrees/disagrees with premise], I want to know whether [premise is correct/incorrect].

Examples:

The majority has been wrong all the time. People used to think that the sun goes around the earth! I don't care how many scientists believe in global warming, I want to know whether the evidence is strong enough to think it's a real threat.

I don't care if everyone on that subreddit thinks it's right, I want to know why YOU think it's right.

Come on, if all your friends jumped off a bridge, would you jump too?

There are, of course, many other things that can go wrong in the course of a conversation. There will be other concise, effective rebuttals, which can be added to this list.

The Hard Work of Translation (Buddhism)

The issue, as it seems to me, is that almost every text you read on Buddhism does not attempt to do the actual work of translation. The first transmission of Buddhism to the west reified a bunch of translations of terms, such as concentration, equanimity, tranquility, mindfulness, suffering, etc. and works since then have mostly stuck to rearranging these words in different combinations and referencing the same metaphors that have been in use since the time of the Buddha. If these authors had true discernment they would realize that the umpteenth text on 'establishing the noble bases of tranquility secluded from sensuous ignorance' or what-have-you aren't helping anyone who didn't already get the message.

At this point I want to say that I think this approach is 'working' for the fraction of the population it is going to work for. If we want to make the practical fruits of Buddhist practice dramatically more accessible to a broader range of humanity we need people to do the hard work of translation to put the Buddha's teachings in forms that will be accessible to various groups of people.

The hard work of translation is to attempt to use language to point your mind at the same distinctions that the original author was trying to point to. Attempts to do this will inevitably fail in lots of ways, but can hopefully communicate enough of the core message that people can piece together the essential causal relations after which, having had direct experience as a result of skillful practice, they can help to improve the translations further.

So, putting my money where my mouth is, I want to try to produce a translation of what I see as the core causal loop that causes progress on the Buddha's path. I'm attempting this because I believe the core causal loop is actually quite small. The Buddha had a tougher task because he had to explain causation, locus of control, and other critical concepts to farmers from scratch.

To begin with, you may think that the purpose of meditation is to eliminate thoughts. But read the Pali Canon and you find a text rife with concepts, schemas, diagnostic methods for various classifications of mental activity, meditation taxonomies, sensory taxonomies, feedback loops etc. Pretending you're already enlightened and that there isn't hard work to do is something the new agers have borrowed from some crappy spiritual schools of various flavors. I refer to people preaching such messages as mindlessness teachers.

To be clear, a decrease in discursive thought, and especially unpleasant mental contents that don't seem to serve any purpose, are one of many pleasant effects of proper practice, but don't really need to be focused on. It is a benefit that arrives in stages on its own.

So, what is the core loop?

It's basically cognitive behavioral therapy, supercharged with a mental state more intense than most pharmaceuticals.

There are two categories of practice, one for cultivating the useful mental state, the other uses that mental state to investigate the causal linkages between various parts

of your perception (physical sensations, emotional tones, and mental reactions) which leads to clearing out of old linkages that weren't constructed well.

You have physical sensations in the course of life. Your nervous system reacts to these sensations with high or low valence (positive, negative, neutral) and arousal (sympathetic and parasympathetic nervous system activation), your mind reacts to these now-emotion-laden sensations with activity (mental image, mental talk) out of which you then build stories to make sense of your situation.

The key insight that drives everything is the knowledge (and later, direct experience) that this system isn't wired up efficiently. Importantly: I don't mean this in a normative way. Like you should wire it the way I say just because, but in the 'this type of circuit only needs 20 nand gates, why are there 60 and why is it shunting excess voltage into the anger circuits over there that have nothing to do with this computation?' way. Regardless of possible arguments over an ultimately 'correct' way to wire everything, there are very low hanging fruit in terms of improvements that will help you effectively pursue *any* other goal you set your mind to.

Funny aside, emotional 'resistance' might be well named, it might be literal electrical resistance in the CNSs wiring as a result of this spaghetti logic.

So back to these stories and story building blocks that are the outputs of this system. You generated a bunch of the primitive building blocks when you were very young and throwing everything together on an as needed basis with no instructions. You both have a back log of such stories and story building-blocks and are generating new ones all the time. Practice improves each of these situations. It improves the backlog by going through and reprocessing stories that aren't actually reality aligned when examined. Again, not pointing to edge cases here but things in the 'your partner humming the spongebob theme shouldn't make you furious because of something that happened when you were 12' class. You can clean up all the obvious stuff and then let your future self (who now has more resources) think about how to wisely deal with the fuzzy edge cases. It improves the new stories coming in (partially by learning as it processes the back log) by building far fewer incoherent stories out of pieces that don't fit together, and building less of the crappier building blocks in the first place.

I'll go ahead and name these things now to connect them up for people who have some knowledge of existing translations.

Concentration meditation gives rise to a mental state where the mind is very calm and inclined to neutrality. Of the same sort you'd want in a good judge.

Insight meditation makes one aware of the causal links in the perceptual system between physical sensations, feelings, and mental reactions.

Sankharas are the stories and story pieces that get reexamined and refactored as a result.

So what is the core loop of meditation practice?

Concentration puts you in the ideal state for insight.

Insight stirs up Sankaras.

Examining Sankharas riles up the mind, eventually leading to a desire to do some more concentration in order to calm down and keep making progress.

Clearing Sankharas cause concentration to go much better. And onward.

Why is concentration ideal to prepare you for insight practice?

Insight requires a high degree of temporal and spatial resolution in order to see the finer linkages between mental activities that normally flow past you without you noticing. Concentration meditation improves that resolution.

Second, to examine the Sankharas is to, to some extent, reactivate the sensations, feelings, and mental reactions associated with them. Since the ones we are most concerned with are the ones that are causing the biggest negative reactions in our lives, we need the mind to be calm and tranquil in order to do this work. Concentration greatly improves this tranquility as well.

How do insights stir up Sankharas?

This would require more speculation about somatic theories that don't yet have a good evidence base. Subjectively, it feels like building up insights into particular kinds of linkages between physical sensations, feelings, and mental reactions causes areas of your backlog that are particularly heavy in those linkages to get some activation and thus be available to consciousness.

You've experienced this if you've ever had a conceptual insight and then spent the next week noticing ways it was applicable, seemingly spontaneously. The only difference here is that insight can also be non-conceptual (ie, insight into how two particular physical sensations interact might generate no verbal content/mental talk but some sense of something happening.)

How does clearing Sankharas improve concentration?

The mental talk, emotional avoidance, and physical discomforts that interrupt concentration practice are built from unendorsed linkages.

So, the Buddha taught a method of concentration, a system for developing insight that we know as mindfulness, and to use these to both stop building new stories and to clear out our backlog of stories. That's actually it. The rest is details for how this plays out in practice. Failure modes can get a bit weird, and even if you do it right some mind blowing states and experiences can pop up. So there's lots of whataboutism for all that.

The miswired central nervous system story gives us simple answers to things like trauma (extreme levels of miswiring of things into fear and freeze responses), why stuff like yoga and exercise help (general CNS health, probably capacitance/fuse breaker improvements), why psychotherapy sometimes but not always activates childhood memories and the significance of that, and why practitioners claim they have a much better life but can't always explain why (they perform the same actions but with much less internal resistance).

So then why all the rest of this crap?

Well, besides my post on why practitioners make so many metaphysical claims, it's also just that there's a lot of idiosyncrasy in first unwiring a randomly wired CNS and then rewiring it in arbitrary order. Especially when you don't really know that that's what you're doing as you're doing it and your mindlessness teacher is a bit clueless as well (though may still have good pragmatic advice despite bad epistemics.)

In addition, each of the practices is actually a practice category. The Buddha taught one specific concentration technique and a simple series of insight techniques, but there are probably a dozen alternatives in each category that seem to work for some people and which entire traditions have subsequently built themselves around and gotten into fights with rival schools about.

Note: I am fairly confident this is how things work up until 2nd path. Since approximately zero percent of people make it beyond that point I'm not too worried about this.

[Answer] Why wasn't science invented in China?

Credit for the question to Eli Tyre.

Preface + Epistemic Status

I spent roughly two days attempting to learn the answer this question plus several more writing it up. What is presented is more accurately described as a *partial answer* or *contribution* towards answering the question - this report isn't actually a confident, solid answer. The question is too large for that.

I need to provide a couple of epistemic caveats:

- I am not experienced at this kind of research, I don't know what kind of rookie mistakes I might be making.
- I have not attempted to assess the reliability of the historians who I quote, though my prior is to be less than completely confident. It might to be too easy to start with a conclusion, a nice narrative, and just find evidence for that. Though I quote what they say, I urge skepticism.
    - I have relatively more trust in statements like "Ibn al-Haytham lived in Cairo in the around the year 1000 and wrote the *Book of Optics*" more than broad statements like "[there was a] virtual absence in ancient Chinese philosophy of anything resembling the socratic method."
- Writing a research report like this one turned out to be remarkably effortful, more so than the pure research and obtaining answers for myself. To keep it from being even longer and more time-consuming, I found myself having to simplify heavily and then often fearing that the result is too simple and leaves too much out (and possibly important stuff out).

Clarifying and Refining the Question

If you don't know why you're doing something, you're likely to instantiate a version of it which doesn't help your upstream goal, assuming it's the right general thing at all. This applies to research too: almost all questions are sub-questions.

In this case, I didn't have an expansion of the question from its author, but I did have their other questions and was aware of their general interests. Other questions they had were: *Are there patterns in what makes great scientists? Were there other intellectual subcultures anything like the rationality community any time in the past few hundred years? What was it that made the intellectual high points of human history high? What made the renaissance, the enlightenment, etc., work?*

The real agenda is clearly about understanding the factors which cause intellectual progress. The expanded question is more like *What were the factors which were present in Europe but absent in China which led to science being invented in the former? In particularly, were there factors which caused Europeans to be more intellectually generative on this front than the Chinese?*

Clarifying terms: Science, Scientific Method, etc.

A few terms worth disentangling: *empiricism* is the notion that you should look at the world to learn about it, and that's a very old idea even when not widely adopted. I wouldn't equate *science* with empiricism for this question. *Science* can either refer to a body of knowledge or the method used by which that knowledge is generated. Though they're tightly connected, I've interpreted this question is primarily about "why wasn't *the scientific method* invented in China?"

Also in this space is the Scientific Revolution and, somewhat more weakly, the Industrial Revolution. The research here has some relevance to those, but I wasn't trying to answer questions about why they didn't occur in China which I expect to possibly involve different economic and social factors then the pure intellectual development of scientific methods by some individuals. I'm looking at invention, not adoption.

Abridged History of the Scientific Method

What follows here is a slightly shortened and heavily simplified summary of Wikipedia entries, primarily from History of the Scientific Method with embellishment from other pages.

Different Scientific Methods over Time

First, it's important to note that there hasn't been just a single *the* scientific method which we can point to having been invented at a single time and space. There have been successively refined methods of generating scientific knowledge developed over time. Scientific methods were possessed by:

- at least one person who wrote an Egyptian medical textbook, (c. 1600 BCE)
- the Babylonians with their mathematical astronomy
- the Greeks (who were foundational)
- the Arabs
- the Chinese Mohists (more on them later)
- the Indian Charvaka school.

This is important because it means in answering the question I'm not looking for factors which caused something to happen at a very particular time and place, e.g. not what made Francis Bacon very special or the like. Instead, I'm looking for factors which held over Europe (and the Middle East) for over a thousand years.

To better understand what the relevant factors might be and to have a more precise of idea of the thing we care about having been invented, I read through the history of scientific methods as developed in Europe.

An ongoing debate in the philosophy of science is the relative role of observation vs reasoning in the scientific method, e.g. *rationalism* vs *empiricism,* and by extension the role and nature of experimentation. Another difference is whether a theory starts with general theories which leads to experimental work or the reverse. The different scientific methods from different thinkers were largely playing with the same elements. Still, they are united by all involving some degree of empiricism, some degree on reason, and are for the purpose of producing naturalistic explanations of physical reality.

In his work on logic, The *Organon* (Greek: Ὄργανον, meaning "instrument, tool, organ") Aristotle (384 BCE to 322 BCE) lays out his *inductive-deductive* method of generating scientific knowledge. Short version: a) via inductions one can discover universals by generalization, b) using deductive reasoning in the form of syllogisms, one can infer new universal truths from those already established. However, Aristotle did not consider the inductive step to be scientific reasoning itself - merely preliminary to real the business which

was the deductive reasoning. It's also worth noting that Aristotle performed no modern-style experiments.

Several scientists from the Middle East pushed much closer to what we recognize as modern science. Arab physicist Ibn al-Haytham (~965 to ~1040) combined observation, experiments, and reasoned argument to his study of optics proving the models held by Ptolemy, Euclid, and Aristotle to be wrong. He criticized Aristotle's handling of induction. Persian scientist Al-Biruni (973 to ~1050) used repeated experimentation in his work and was concerned with systematic errors and observational biases. He held that "universals came out of experimental work" and "theories are formulated after discoveries." Ibn Sina (Avicenna, 980 to 1037) combined induction and experimentation, criticizing Aristotle's induction with the claim "it does not lead to the absolute, universal, and certain premises that it purports to provide." Ibn Sina might also have been the first to describe several of the methods of induction listed by John Stuart Mill in 1843. Avicenna's scientific method is one in which "general and universal questions came first and led to experimental work." [Note these guys all seemed very impressive to me.]

During the Renaissance of the 12th century, ideas on scientific methodology, including Aristotle's empiricism and the experimental approaches of Ibn al-Haythem and Ibn Sina were introduced to medieval Europe via Latin translations of Arabic and Greek texts together with commentaries.

Robert Grosseteste (~1175 to 1253) was probably one of the first European thinkers in Europe to understand Aristotle's vision of the dual nature of scientific reasoning. His conception was of going from particular observations to universal laws back to prediction of particular observations. "Resolution and composition." He's been called the real founder of the tradition of scientific thought in medieval Oxford.

Roger Bacon (~1219 to 1292) was inspired by the writings of Grosseteste. He describes a repeating cycle of observation, hypothesis, experimentation plus the need for independent verification. With special permission from the Pope (necessary since he was a friar) he published three large treatises.

At this point in the history it is noted that in 1562 "Outlines of Pyrrhonism" by Sextus Empiricus (c. 160) were printed in Latin and circulated in Europe, placing the arguments of classical skepticism in the European mainstream.

In 1620, Francis Bacon (1561 to 1626) published his *Novum Organum,* the title a reference to Aristotle's *Organon*. Countering Aristotle, he said that induction must be used "for proving and discovering not first principles only, but also lesser axioms, the middle, and indeed all." Unlike Aristotle, Bacon insists on induction throughout the entire process, not just at the beginning to derive universals. He was very committed to experimentation, including "crucial experiments" to differentiate between competing hypotheses. However, unlike in modern scientific process, hypothesizing forms only a small part of Bacon's method. In his method hypotheses were supposed to arise in the process of investigation (contrast this with Avicenna who was happy to start with general theories and then experiment).

Isaac Newton (1642 to ~1726) embraced Bacon's empiricism and outlined *four rules on reasoning* in the Principia. Some of his methods were systematized by John Stuart Mill.

The "hypothetico-deductive" method with its focus on the formulation and testing of hypotheses which can be falsified arose in the 18th century. Major contributors towards this refinement of the method were Charles Sanders Pierce (1839 to 1914) and Karl Popper (1902 to 1994). Pierce made induction and deduction complementary and put forth the basic schema for hypothesis testing we have until this day. Popper is the famed champion of falsification.

In the 20th century, Bayes' theorem was brought to bear on the scientific method, though it is a lens not yet universally adopted.

So when was “science invented”?

When starting this project, my vague conception was something like “the scientific method was invented in the 17th century by Francis Bacon.” After reading this history, I realize what an extreme, perhaps laughable, simplification - the things you believe when you haven’t thought about them for even five minutes.

The modern scientific method as we know it is the result of over a thousand years of intellectual tradition. It was built piece by piece, eminent scholars each providing their own contribution. Contrary to my starting conception, it’s not that meaningful to say that it was “invented” specifically in the 17th century.

However, it does still seem that something significant happened in the 17th century or then abouts. I didn’t go into this in any depth, but the following pieces might be involved:

- The improvements in the methods Francis Bacon and Newton caused a dramatic difference in the power of the scientific method leading to significant breakthroughs.
- Major scientific breakthroughs, caused by improved method or other causes, occurred around this time causing widespread interest and enthusiasm for science, leading to social change. The social change cause these developments to be especially noteworthy.
- Social change happening for multiple reasons at this time caused the improved scientific method to gain a lot of traction at this time.
- Newton and Galileo (though not Bacon) had impressive scientific breakthroughs which caused their methods to get a lot of attention.

I’ll quote the entry on Scientific Revolution:

> In the 19th century, William Whewell described the revolution in science itself—the scientific method—that had taken place in the 15th–16th century. "Among the most conspicuous of the revolutions which opinions on this subject have undergone, is the transition from an implicit trust in the internal powers of man's mind to a professed dependence upon external observation; and from an unbounded reverence for the wisdom of the past, to a fervid expectation of change and improvement."[12] This gave rise to the common view of the Scientific Revolution today:
>
> A new view of nature emerged, replacing the Greek view that had dominated science for almost 2,000 years. Science became an autonomous discipline, distinct from both philosophy and technology and came to be regarded as having utilitarian goals.[13]

. . .

> In 1984, Joseph Ben-David wrote:
>
> *Rapid accumulation of knowledge, which has characterized the development of science since the 17th century, had never occurred before that time.* ***The new kind of scientific activity emerged only in a few countries of Western Europe, and it was restricted to that small area for about two hundred years.***

This last quote raises an interested related question to the main one being answered here: if true, what factors caused the rapid accumulation of knowledge in specifically only a few countries and for only those two hundred years?

Perhaps the issue was that China wasn’t included in the intellectual tradition?

Reading the history of the scientific method and noting that there was this long intellectual tradition and it depended on translations and ideas spreading from one place to another, I wondered whether the lack of (supposed) scientific development in China was that China wasn't exposed to this tradition. Perhaps they didn't get the translations. Later reading suggested this wasn't true -- China had sufficient exposure but for some reason didn't latch on and didn't develop its own science or scientific method in the same way.

Toby Huff writes in *The Rise of Early Modern Science: Islam, China, and the West, 2nd Edition*:

> By the end of the fourteenth century in the areas of mathematics, astronomy, and optics, there was a considerable debit on the Chinese side, **despite the fact that there had been many chances for the Chinese to benefit from Arab astronomers and to borrow or assimilate the Greek philosophical heritage through constant interchanges between the Arabs and the Chinese.** During Yuan times (ca. 1264-1368) Needham tells us, the Arabs or, more probably, the Persians played a significant role in bringing new mathematical ideas to Chinese science, and this role paralleled that played by Indians in T'ang times. (p. 242)

Apparently the Chinese ended up employing Muslim astronomers in their astronomy bureau because they hadn't mastered their superior methods yet.

> Their deficiencies in this area, moreover, led the Chinese to employ Muslim astronomers in the Chinese Bureau of Astronomy continuously from the thirteenth century onward. Indeed, in 1368 a special Muslim Bureau of Astronomy was established in China that was still functioning at the time of the arrival of the Jesuits in the sixteenth century.16 Upon the arrival of the Jesuits, there were four competing astronomical systems: the traditional Chinese system; that of the Muslims (based on the lunar calendar); the new European; and that of the so-called new Eastern Bureau.17 For these reasons Needham notes that "there can be no doubt but that there was every opportunity for Arabic and Persian mathematical influences (as from the observations of Maragha and Samarqand) to enter Chinese traditions."18 Even more tantalizing are the reports that a Mongol ruler in China, Mangu (d. 1257; the brother of Hulagu who ordered the construction of the Maragha observatory), is said to "have mastered difficult passages of Euclid by himself."19 In what language was this version of Euclid, and why is it that Mangu's successor - Khubilai Khan - did not suggest the learning of Euclid to the court officials surrounding him ?20 These facts make it all the more puzzling why it was that the Jesuits are credited with having introduced Western astronomy to the Chinese (albeit incompletely because of the Galilean controversy just then unfolding) as well as geometry, when the Maragha models clearly assumed all the fundamentals of Western astronomy at that time except the heliocentric orientation. In other words, given the direct contact in the capital city between some of the best Muslim astronomers of the time and the Chinese astronomers in the official Bureau of Astronomy, the Chinese ought to have had nearly two centuries to translate Euclid's Elements and to assimilate the Ptolemaic models (as perfected by al-Tusi, al-'Urdi, al-Shirazi, and Ibn al-Shatir) before they were transformed into the Copernican models by Europeans in the sixteenth and seventeenth centuries. (p. 244)

It looks like the Chinese had many opportunities to absorb and build upon foreign knowledge. For some reason they didn't, instead employing foreigners for centuries. This is evidence against my hypothesis that China wasn't as intellectual generative because they didn't get to be part of the same intellectual tradition. They could have been, but something got in the way.

Perhaps China didn't have networks of scholars the same way?

Related to the idea of an intellectual tradition which built upon itself, I believe I hit upon the topic of various communities of thinkers which exchanged ideas in Europe.

Invisible College is the term used for a small community of interacting scholars who often met face-to-face, exchanged ideas, and encouraged each other. One is examples is Robert Boyle's network of natural philosophers who had a focus on acquiring knowledge through experimentation. They were supposedly a precursor to the Royal Society. The Hartlib Circle, a network of correspondence across Western and Central Europe, was another instance of an invisible college.

Relatedly, the Republic of Letters was a long-distance intellectual community in the late 17th and 18th centuries in Europe and the Americas. I didn't look into this or any of the others, but the Republic of Letters was bound up with the Royal Society, famed institute and network of Science.

It's been asserted [source] that having Latin as a lingua franca was important for Europe *integrated market for ideas* . Makes sense if scholars who otherwise speak different languages are going to be able to communicate.

These communities and networks of thinkers maybe a factor which contributed towards the development of the scientific, its spread, and associated social change. A question I had, yet didn't fully get into, is the extent to which China has similar communities and networks of thinkers. I do have some early indication that travel was more difficult in China and that China had a powerful censoring system, though it is unclear to me the effect that censor would have had on scholars exchanging letters given that its primary target were officials in the bureaucracy.

Science and Scientific Methods in China

Main article: *History of Science and Technology in China*

An Existing Literature

Working on this project, I found that there is an existing body of work on the related questions *"why wasn't science invented in China?", "why didn't China have a scientific revolution?", "why didn't China have an industrial revolution?" and "why did China fall behind Europe?"*

A launching for this literature is the Scientific and technological stagnation section of the History of science and technology in China Wikipedia entry. And relatedly, The Great Divergence

Key historians on this topic include Joseph Needham, Toby E. Huff, Nathan Sivin, Derek Bodde and Justin Lin. Most of the early Western work in the history of science in China was done by Needham. He wrote a series of books/encyclopedia called *Science and Civilisation in China*. In 1969 he asked:

> "Why did modern science, the mathematization of hypotheses about Nature, with all its implications for advanced technology, take its meteoric rise only in the West at the time of Galileo?" "Why modern science had not developed in Chinese civilization … ?" He adds a second question that makes the larger problem more interesting: "why, between the first century B.C. and the fifteenth century A.D., Chinese civilization was much more efficient than occidental in applying human natural knowledge to practical human needs." (Needham cited by Sivin in Why the Scientific Revolution Did Not Take Place in China --or Didn't It? P. 2)

Is it a reasonable question?

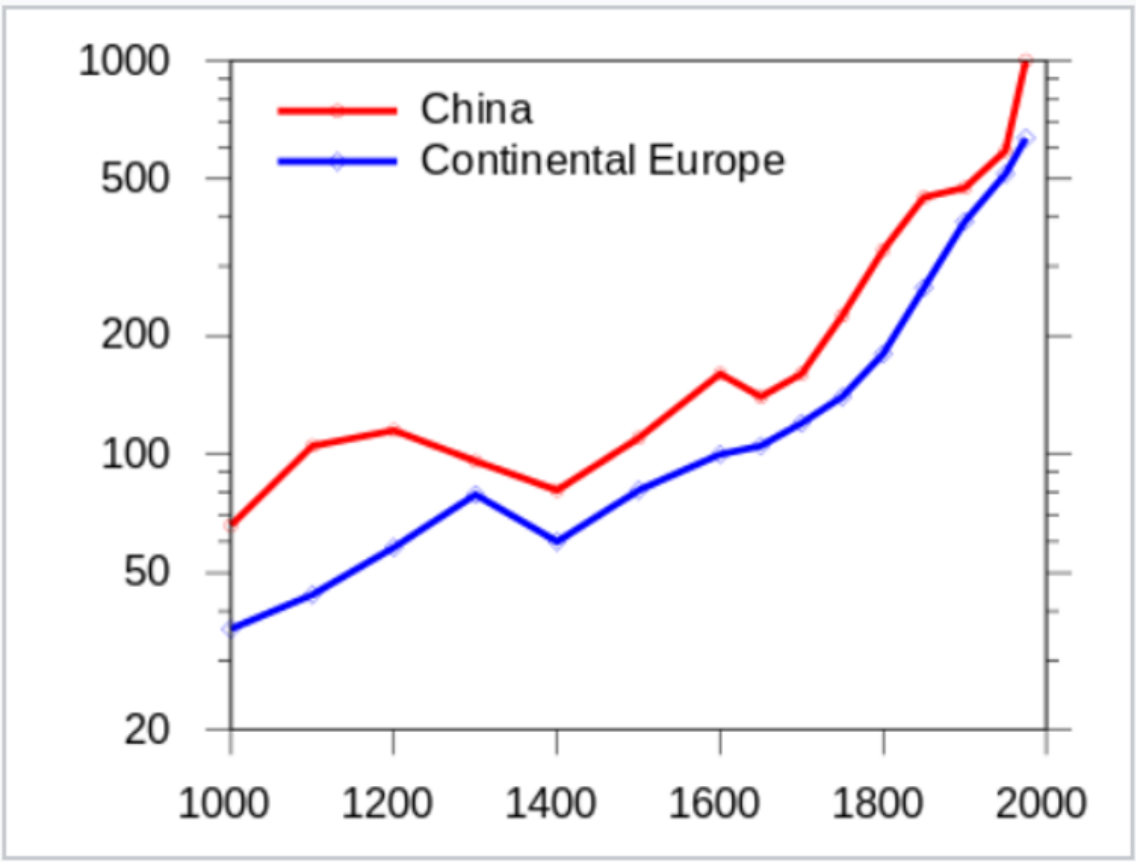
At least some question that asking why science/the scientific revolution didn't happen in China is a good question. Maybe it's like asking why your neighbor's house didn't burn down. But there are some good reasons.

First, China had a long history of being quite technologically advanced.

> "Chinese civilization was much **more** efficient than occidental [civilization] in applying human natural knowledge to practical human needs." (Needham cited by Huff, p. 241)

They invented the abacus, the crossbow and the *Four Great Inventions* of the compass, gunpowder, papermaking, and printing which Francis Bacon considered the most important inventions facilitating the West's transformation from Dark Ages to the Modern world [Lin]. They possessed these a thousand years before Europe did. Early China had matches, dry docks, piston pumps, cast iron, the wheelbarrow, parachute, natural gas as fuel, and the suspension bridge. They had technology and they had a wide range of sciences. They had many sciences even if they didn't have a unified science [source].

Second, China had a popular large than Continental Europe [source].



Comparative populations (millions, log scale) of China and Continental Europe between 1000 and 1975.[36]

Given technological prowess and population size, it doesn't seem unreasonable to me to ask why China didn't generate the modern scientific method or undergo a scientific revolution.

So why not in China?

I collect here reasons from Toby E. Huff's book *The Rise of Early Modern Science* (overwhelmingly), Wikipedia, and a few other papers I glanced at. Huff's book itself is something of a summary which is incorporating views from Needham, Sivin, Bodde, and others. What is presented here is primarily taken from Chapters 6 and 7 from Huff's book, about 75 pages. It is a very interesting and worthwhile read which I've done a so-so job of summarizing. Possibly this answer could have been framed as a book review.

Some of reasons taken from the Wikipedia entry on *The Great Divergence* [between China and the West]. These are less directly about the development of science specifically than overall economic divergence, but I've included the few which might also apply in the specific case here.

Huff writes:

> If one takes the point of view that science is above all a system of error detection, not a set of skills for building machines, mechanical or electronic, **then attention must be directed toward those abstract systems of thought and explanation that give higher order to our thinking about the natural realm.** Science at its heart is systematic and theoretical knowledge about how the world is and how it works. It is episteme as opposed to techne. It is speculative in that it is always conjecturing the existence of new entities, processes, and mechanisms, not to mention possible new worlds. Its task is to determine which of these ideas and entities have a real existence in the world. Karl Popper's description of this process as "conjectures and refutations" aptly captures this dynamic.4 From such a point of view, science is about how to describe, explain, and think about the world and is not concerned with how to make labor easier or how to control nature. (p. 241)

We are differentially looking for factors which facilitate the develop of science and scientific thinking, not technological progress more broadly. Building technology is a safer endeavor, politically and metaphysically, with less opportunity to challenge the prevailing worldview.

Overview of Huff

Huff's overall thesis is that China was unable to produce modern science primarily because a lack of the requisite intellectual freedom. Unlike Europe, China's political, religious, legal, and education systems did not afford the *neutral spaces* where novel ideas could be advanced and old ideas questions. Culturally, it may have lacked the practice of debate and dispute, including anything resembling the Socratic dialog. Further, philosophically and conceptually, China's neo-Confucian worldview which viewed the world through correlations and binary pairs may not have lent it itself to the causal thinking necessary for science.

> . . . it did not encourage or tolerate thinkers who were essentially disputatious and critical of the intellectual status quo . . There was no Chinese equivalent to the Scholastic method of disputation, no canons of logic a la Aristotle, and no mathematical methods of proof such as one finds in Euclid's geometry. Derk Bodde points out, "Throughout its history Confucianism has deprecated the use of debate as a means of advancing knowledge." (p. 279)

I expand on these points in the following sections. Unless otherwise stated, statements are coming sourced from Huff.

More Detailed Summary of Huff

In the 12th century, Europe had a legal revolution which redefined the nature of social organization with significant political, social, and economic, and religious effects. In particular, legally autonomous collectives emerged including cities, towns, universities, interest groups, and professional groups. While overall subject to an overall law, these groups could still operate fairly autonomously, and could set their own rules and regulations, could owe land, could sue, could have legal representation before the king's court, and generally operate without too much interference from the authorities.

I didn't properly read or think about Huff's chapters on the legal revolution and its significance, and therefore can't do it justice, however other relevant points: a) this being the beginning of separation of state and religion, b) this legal revolution represented an adaptation of Roman civil to the European's needs.

[Huff makes the broad assertion that modern science arose in Europe due to the unique combination of Greek philosophy, Roman law, and Christian theology. I didn't read enough to

fully understand this model, however.]

Legally autonomous collectives of special interest to the development of science are the European universities which arose in 12th century around the time of the legal revolution. Though many grew out of cathedral schools and religious orders, that wasn't always the case, and the universities "sprang into existence . . . without express authorization of king, pope, prince, or prelate." Crucially, they were able to set their own curricula. Practically, they taught a lot of Greek philosophy including its naturalistic and scientific parts.

> In the Byzantine Empire, the "pagan curriculum" of the Greeks, based on the seven liberal arts (grammar, rhetoric, logic, followed by arithmetic, geometry, harmonics, and astronomy [the quadrivium]), was taught from the "fifth to the fifteenth century."3 Naturally, the curriculum was infused with the Greek philosophical outlook, and that doubtless was the rub. (p. 140)

In contrast to the legal autonomy developing in Europe, from the 12th century (and earlier) China was (with only a few exceptions) a unified, top-down and bureaucratically run empire where there was no separation of state and religion and all power flowed down from the emperor upon whom there were no legal checks or balances.

Legally and bureaucratically, there was a hierarchy of officials which eventually rolled up into the emperor. Lower-level officials had almost no authority or independence from higher ups. Any collectives which might be created were not autonomous from the larger empire. There was nowhere safe to venture any revolt against authority.

The Chinese also did not have a system of law as we would recognize in the west (a system with universally applicable rules, rights, due process, interpretation of the law) and instead something more simply a penal code with no clear rules, variable enforcement, and many exceptions to be made at the discretion of whichever authority is applying punishment.

The legal system is relevant is both to lack of autonomous spaces and also the possibility that due to lack of a conception of a universal law which was binding to all citizens, the Chinese lacked the metaphor to laws which were binding to nature, and therefore did not seek them out.

Significantly, the vast Chinese imperial bureaucracy was staffed semi-meritocratically based on the imperial examination. This system stood intact from 1400 to 1905. The effect of the examination was standardize Chinese education. However, it was rigidly controlled and focused on literary and moral learning - not mathematics or the sciences - and largely consisted of rote memorization of hundreds of thousands of words of Chinese classic. There are nuances to these states as at various periods there were questions about mathematica, music, and astronomy on the exam, however they were never a focus. In contrast, European universities were free to set their own curricula.

Philosophically, in place of Western atomism/reductionism governed by mechanical and impersonal laws of nature, China viewed the world as made of primary forces (yin and yang) and the five phases (metal, wood, water, fire, and earth) constantly shifting in recurrent cycles. Things are not to be understood through laws governing parts, but through the unity of the whole. The patterns of the natural world were studied to find correlative correspondences between the conduct of the emperor and the patterns of the heavens. In this vein, there was a strong focus on binary pairs like light vs dark, hot vs cold, heaven vs earth - but all understand as natural complements which follow an inscrutable patterned progression. To translate into more typical LessWrong language, the Chinese didn't look for gears-levels models, because they didn't have a conception of gears - just patterns in the organic unity of the whole. Or something.

Even if we grant Aristotle was a little confused, we can see he was trying to explain things in terms of universal laws which explained things with a causal structure. From that point, you can progress to finding better universal explanations. I can imagine if you're not thinking

causally, you can’t embark on a process of science. Though I must confess, having very little grasp of Chinese philosophy, I cannot appraise the claims about historical Chinese philosophy and worldview.

An exception of the above point on philosophy is that the Chinese Mohist school of thought might have been on track to develop a solid scientific method, but unfortunately they and their thinking died out.

> Although it seems doubtful that early Chinese methodological discussions were equivalent to those of Aristotle and Plato, it must be said that in the work of Mo-tzu (fourth century B.C.) there are keen methodological insights that, in Needham's words, "could have become the fundamental basic conceptions of natural science in Asia. "38 Perhaps one could even agree with Needham that the Mohists "sketched out what amounts to a complete theory of scientific method. "39 The problem is that the Mohists and their thought faded into Chinese history and apparently had little influence on Chinese natural thinkers and none at all on Western thought. Despite the promising beginnings one sees in Mohist philosophical thought, it never gained much influence in the Chinese thought world. (p. 247)

Other Factors

There are at least a couple more factors not covered in Huff’s work which contributed to the rise of modern science Europe. rather than China.

Political fragmentation

It may be macro-politically relevant that because China was a unified empire, there was nowhere that thinkers seeking more intellectual freedom could migrate too. In Europe, if a nation’s policies hampered scientific or technological progress they would be soon be outcompeted by others who didn’t. Jared Diamond, author of *Guns, Germs, and Steel* that the European balkanization into smaller states was the result of geography with China’s geography being more conducive to a large, monolithic, isolated empire. [Wikipedia]

Economic freedom

Related to intellectual freedom, economic freedom may have been quite relevant. David S. Landes writes in *Why Europe and the West? Why not China?*:

> China had two chances: first, to generate a continuing, self-sustaining process of scientific and technological advance on the basis of its indigenous traditions and achievements; and second, to learn from European science and technology once the foreign "barbarians" entered the Chinese domain in the sixteenth century. China failed both times. What explains the first failure? **I stress the role of the market: the fact that enterprise was free in Europe while China lacked a free market and institutionalized property rights; that in Europe innovation worked and paid, while the Chinese state was always stepping in to interfere with private enterprise.** As for the second failure, China's cultural triumphalism combined with petty downward tyranny made it a singularly bad learner.

I did not explore this factor. The natural question to me here is how much the potential for profit or other gain incentivized the development of science. For example, astronomy was arguably the foundational scientific field and had important applications. How did these incentives affect academics, scholars, and the men of means who already had enough wealth they could engage in scientific discovery?

Europe’s escape from the Malthusian trap

My colleague, jimrandomh, brought this to my attention. The thesis is that due to plague, war, and urbanization, early Modern Europe escaped the Malthusian trap which kept everyone at subsistence with no resources left for advancing science and technology. One source for this is here.

The High-level equilibrium trap

A similar economic point has been made by Mark Elvin to explain why China didn't develop its own revolution. From the Wikipedia entry:

> Essentially, he [Mark Elvin] claims that the Chinese pre-industrial economy had reached an equilibrium point where supply and demand were well-balanced. Late imperial production methods and trade networks were so efficient and labor was so cheap that investment in capital to improve efficiency would not be profitable.

The relevance to the development, as above, would depend on how strongly the demand for profit-increasing technology impacted the development of science.

See *The Great Divergence* for more factors and more detail.

Conclusion

This project simultaneously took far more time than expected and yet also still feels very shallow, heavily simplified, and bluntly presented. I do find Huff's account, which I have cited extensively quite compelling, but I have made no attempt to verify his reliability or the reliability of his sources.

Though it is already intuitive to say that intellectual freedom was crucial for the development of science, I think Huff's work is an impressively detailed case that Europe had this freedom in a way China did not. I did not cover it here, but Huff's book also explores the situation in the Islamic world adding another point of comparison to his analysis.

My LessWrong associations are biasing me here, but it does seem important that Europe had a sustained intellectual tradition reaching back through the Arab scholars to the Greeks. The modern scientific method wasn't developed de novo, but rather it was built up by piece by piece from antiquity. I think that's an argument for technologies that better collect, curate, and transmit knowledge.

I didn't explore economic explanations more, and there might be something to them, but I imagine that the political, religious, and cultural structures upstream of intellectual freedom were also upstream of economic factors. A fully comprehensive picture would probably include all three.

Overall, I do conclude a firm conclusion that even I'm not sure of the details, there were almost certainly concrete systematic factors which caused Europe to develop modern science and the scientific method even when China and the Islamic world didn't. And that studying these factors probably does help us identify which factors are core to further intellectual progress in the present.

Related Questions

- What made the ancient Greeks so generative? It seems they founded the Western philosophical and scientific traditions, but what led to their innovativeness?
- What is the relationship between the development of science and economic incentives to do so?

- Q. What kind of things where the people who innovated on science researching? E.g. those from history of science. How much immediate practical value did those things have?
- What is a detailed account of how Bacon's scientific method different from Aristotle's inductive-deductive method?
- How much was the scientific revolution caused by the innovations in the scientific methods at the time vs more general circumstantial factors?
- How much was science/the scientific method responsible for the Great Divergence?
- Q. What did China have which was similar to the Republic of Letters, Royal Society, Invisible College?
- Q. Why exactly was astronomy so important? Time-keeping, setting calendars, navigation . . . I'm interested in a more detailed understanding of how astronomical knowledge was used and how much it mattered.
- Is Joseph Ben David's claim that there was a rapid accumulation of scientific knowledge in only a few countries of Western Europe and only for about 200 years accurate? If so, why?

Reflections on the Research Experience

My background motivation for this work was to research the research experience as part of the validation and design of LessWrong's efforts to build an Open Questions research platform. I collect here a few observations/notes:

- I did this work with relatively little evaluation of the ideas presented. It seems it would be an additional skill to learn how to evaluate historians and their historical work. Ideally I'd have been thinking quantitatively and assigning credences to the points I reported.
- I do feel somewhat suspicious of Wikipedia, it paints very neat narratives and I'm doubtful that reality is that neat.
- In our early thinking, the LessWrong team was approaching research with something roughly like asking a question, asking sub-question, asking sub-sub-questions and so on. There are elements of this, but overall it didn't feel like that. A lot of it felt like realizing I needed knowledge of a topic and then just trying become a mini-expert in that topic. For example, having decided I need to understand the development of the scientific method, I was just reading a bunch of Wiki articles not holding explicit questions in my mind, instead letting my intuition guide me.
- Writing is time-consuming and constituted the bulk of time spent on this project.
    - When reading just for myself, it was okay if not every idea was crystallized, however to write things for others I really had to first get them clear in my own head.
    - I'm not sure how the writing effort scales with the amount of research done, but I'm pretty sure it's sub-linear. More research might actually result in a much clearer picture making it easier to write.
        - During the writing stage, I was progressively reading more and more of Huff's book which actually made it easier to write and I more properly absorbed the picture he was painting.
- I wonder about the alternative approaches to conducting and communicating this research. Possibly I went into too much depth (as shallow as it felt) and would have been better to write much more condensed, lossy summaries. And either separately or additionally, perhaps provide an annotated set of references, i.e. collecting all the sources I read, explaining what one should read to arrive that the same picture I developed.
- I wonder how much this research could have been split into parts. Was it necessary that same person researched both the history of the scientific method and also the history of science in China? It feels like I had a better perspective for having done so, but maybe it wasn't necessary and this could have been two smaller projects.

What are some good examples of incorrigibility?

The idea of corrigibility is roughly that an AI should be aware that it may have faults, and therefore allow and facilitate human operators to correct these faults. I'm especially interested in scenarios where the AI system controls a particular input channel that is supposed to be used to control it, such as a shutdown button, a switch used to alter its mode of operation, or another device used to control its motivation. I'm especially interested in examples that would be compelling within the machine learning discipline, so I'd favour experimental accounts over hypotheticals.

LW Update 2019-04-02 – Frontpage Rework

Since LW2.0 launched, the frontpage had become very complex – both visually and conceptually. This was producing an overall bad experience, and making it hard for the team to add or scale up features (such as Q&A, and later on Community, Library and upcoming Recommendations)

For the past couple months, we've been working on an overhaul of the frontpage (and correspondingly, the overall site design). Our goal was is to rearrange that complexity, spending fewer "complexity points" on things that didn't need them as much, so we could spend them elsewhere.

Frontpage Updates

- Tooltip oriented design.
    - It's easier to figure out what most things will do before you click on it.
- Navigation Menu
    - Helps establish the overall site hierarchy
    - Available on all major site pages (not Post Pages, where we want people to read without distraction)
    - Improved mobile navigation (shows up as a tab menu at the bottom)
    - Eventually we'll deprecate the old Nav Menu (still available in the header) and replace it with a collapsible version of the new one.
- Home Page streamlining
    - Moved Recommend Sequences and Community over to the Nav Menu, so there are only 3 sections to parse
    - Post Items simplified down to one line.
    - Latest Posts now only have a single setting: "*show personal blogposts*", instead of forcing you to figure out immediately what "meta", "curated" and "daily" are.
    - Post List options are generally 'light cobalt blue' – not too obtrusive, but easier to find when you want them.
- Questions Page now has two sections:
    - *Recent Activity* – simply sorted by "most recently commented at", so if you respond to an old question it will appear above the fold.
    - *Top Questions – also* sorted by "recently commented", but filtered to questions with 40 or more karma, so that it's easier to catch up on updates to highly upvoted questions.
- *Community Page*
    - UI updated to match Home Page.
    - The group section now shows 7 groups instead of 3, and has a load more button.

"Everything is Correlated": An Anthology of the Psychology Debate

This is a linkpost for https://www.gwern.net/Everything

Pecking Order and Flight Leadership

It was recently pointed out to me that humans are weird, compared to other social animals, in that we conflate the *pecking order* with the *group decision-making process*.

The pecking order, for instance in birds, is literally the ranking of who gets to eat first when food is scarce.

We can also call it a "dominance hierarchy", but the words "dominance" and "hierarchy" call up associations with *human* governance systems like aristocracy and monarchy, where the king or chief is *both* the decisionmaker for the group *and* the person entitled to the most abundant resources.

In birds, it's not like that. Being top chicken doesn't come with the job of "leading" the other chickens anywhere; it just entitles you to eat better (or have better access to other desirable resources). In fact, group decisionmaking (like deciding when and where to migrate) *does* occur in birds, but *not* necessarily according to the "pecking order". Leadership (setting the direction of the group) and dominance (being high in the pecking order) are completely independent in pigeons, for instance. Pigeons have stable, transitive hierarchies of flight leadership, *and* they have stable pecking order hierarchies, and these hierarchies do not correlate.

Logically, it isn't necessary for the individual who *decides what others shall do* to also be the individual who *gets the most goodies*. They *can* be related — one of the things you can do with the power to give instructions is to instruct others to give you more goodies. But you can, at least with nonhuman animals, separate pecking-order hierarchies from decision-making hierarchies.

You can even set this up as a 2×2:

High rank in pecking order, high decision-making power: Lord

High rank in pecking order, low decision-making power: Eloi

Low rank in pecking order, high decision-making power: Morlock

Low rank in pecking order, low decision-making power: Vassal

"Eloi" and "Morlocks" are, of course, borrowed from H.G. Wells' *The Time Machine*, which depicted a human species divided between the privileged, childlike Eloi, and the monstrous underground Morlocks, who farm them for food. Eloi enjoy but don't decide; Morlocks decide but don't enjoy.

The other archetypal example of someone with low rank in the pecking order but high decision-making power is the *prophet.* Biblical prophets told people what to do — they could even give instructions to the king — but they did not enjoy positions of privilege, palaces, many wives, hereditary lands, or anything like that. They *did* sometimes have the power to threaten or punish, which is a sort of "executive" power, but *not* the power to personally enjoy more resources than others.

In American common parlance, "leadership" or "dominance" generally means *both* being at the top of a pecking order *and* being a decision-maker for the

group. My intuition and experience says that if somebody wants to be the decision-maker for the group but *doesn't* seem to be conspicuously seeking & enjoying goodies in zero-sum contexts — in other words, if somebody behaves like a Morlock or prophet — they will read as not behaving like a "leader", and will fail to get a certain kind of emotional trust and buy-in and active participation from others.

My previous post on hierarchy conflated pecking-order hierarchies with decision-making hierarchies. I said that people-telling-others-what-to-do (decision-making hierarchy) "usually goes along with" special privileges or luxuries for the superiors (pecking-order hierarchy.) But, in fact, they *are* different things, and the distinction matters.

Most of the practical advantages of hierarchy in organizations come from decision-making hierarchy. A tree structure, or chain of command, helps get decisions made more efficiently than many-to-many deliberative assemblies. Many of the *inefficiencies* of hierarchy in organizations (expensive displays of deference, poor communication across power distance) are more about pecking-order hierarchy. "So just have decision-making hierarchy without pecking-order hierarchy!" But that's rule-by-prophets, and *in practice people seem to HATE prophets*.

The other model for leadership is the "good king", of the kind that Siderea writes about in this series of posts on Watership Down. The good king is *not* just sitting on top of the pecking order enjoying luxury at the expense of his people. He listens to his people and empowers them to do their best; he shares their privations; he is genuinely committed to the common good. But he's still a king, not a prophet. (In *Watership Down*, there actually *is* a prophet — Fiver — and Hazel, the king, is notable for *listening* to Fiver, while bad leaders ignore their prophets.)

My guess is that the "good king" *does* sit on top of a pecking-order hierarchy, but a very mild and public-spirited one. He's *generous*, as opposed to greedy; but generosity implies that he *could* be greedy if he wanted to. He shares credit with others who do good work, instead of hogging all the credit for himself; but *being the one to give credit* itself makes him seem central and powerful.

A "good king" seems more emotionally sustainable for humans than just having a "prophet", but it could be that there's a way to implement pigeon-like parallel hierarchies for resource-enjoyment and decision-making, or other structures I haven't thought of yet.

Alignment Newsletter One Year Retrospective

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

On April 9, 2018, the first Alignment Newsletter was sent out to me and one test recipient. A year later, it has 889 subscribers and two additional content writers, and is the thing for which I'm best known. In this post I look at the impact of the newsletter and try to figure out what, if anything, should be changed in the future.

(If you don't know about the newsletter, you can learn about it and/or sign up here.)

Summary

In which I badger you to take the 3-minute survey, and summarize some key points.

Actions I'd like you to take

- If you have read at least one issue of the newsletter in the last two months, **take the 3-minute survey!** If you're going to read this post anyway, I'd prefer you *first* read the post and then take the survey; but it's *much* better to take the survey without reading this post than to not take it at all.
- Bookmark or otherwise make sure to know about the spreadsheet of papers, which includes everything sent in the newsletter, and a few other papers as well.
- Now that the newsletter is available in Mandarin (thanks Xiaohu!), I'd be excited to see the newsletter spread to AI researchers in China.
- Give me feedback in the comments so that I can make the newsletter better! I've listed particular topics that I want input on at the end of the post (before the appendix).

Everything else

- The number of subscribers dwarfs the number of people working in AI safety. I'm not sure who the other subscribers are, or what value they get from the newsletter.
- The main benefits of the newsletter are: helping technical researchers keep up with the field, helping junior researchers skill up without mentorship, and reputational effects. The first of these is both the most important one, and the most uncertain one.
- I spent a counterfactual 300-400 hours on the newsletter over the last year.
- Still, in expectation the newsletter seems well worth the time cost, but due to the high uncertainty on the benefits to researchers, it's plausible that the newsletter is not worthwhile.
- There are a bunch of questions I'd like feedback on. Most notably, I want to get a better model of how the newsletter adds value to technical safety researchers.

Newsletter updates

In which I tell you about features of the newsletter that you probably didn't know about.

Spreadsheet

Many of you probably know me as the guy who summarizes a bunch of papers every week. I claim you should instead think of me as the guy who maintains a giant spreadsheet of alignment-related papers, and incidentally also sends out a changelog of the spreadsheet every week. You could use the spreadsheet by reading the changelog every week, but you could also use it in other ways:

- Whenever you want to do a literature review, you find the relevant categories in the spreadsheet and use the summaries to decide which of the papers to read in full.
- When you come across a new, interesting paper, you first Ctrl+F for it in the spreadsheet and read the summary and opinion if they are present, before deciding whether to read the paper in full. I expect most summaries to be more useful for this purpose than reading the abstract; the longer summaries can be more useful than reading the abstract, introduction and conclusion. Perhaps you should do it right now, with (say) "Prosaic AI alignment", just to intuitively get how trivial it is to do.
- When you find an interesting idea or concept, search for related words in the spreadsheet to find other writing on the topic. (This is most useful for non-academic ideas -- for academic ones, Google Scholar is the way to go.)

I find myself using the spreadsheet a couple of times a week, often to remind me of what I thought about a paper or post that I had read a long time ago, but also for literature reviews and finding papers that I vaguely remember that are relevant to what I'm currently thinking about. Of course, I have a better grasp of the spreadsheet making search easy; the categories make intuitive sense to me; and I read far more than the typical researcher, so I'd expect it to significantly more useful to me than to other people. (On the other hand, I don't benefit from discovering new material in the spreadsheet, since I'm usually the one who put it there.)

Translation

Xiaohu Zhu has offered to translate the Alignment Newsletter to Mandarin! His translations can be found here; I also copy them over to the main Alignment Newsletter page. I'd be excited to see more Chinese AI researchers reading the newsletter content.

Newsletter stats

In which I present raw data and questions of uncertainty. This might be useful to understand newsletters broadly, but I won't be drawing any big conclusions. The main takeaway is that lots of people read the newsletter; in particular, there are more subscribers than researchers in the field. Knowing that, you can skip ahead to "Impact of the newsletter" and things should still make sense.

Growth

As of Friday April 5, according to Mailchimp, there are 889 subscribers to the newsletter. Typically, the open rate is just over 50%, and the click-through rate is 10-15%. My understanding is that this is very high relative to other online mailing lists; but that could be because of online shopping mailing lists, where you are incentivized to send lots of emails at the expense of open and click-through rates. There are probably also readers who read the newsletter on the Alignment Forum, LessWrong, or Twitter.

The newsletter typically gets a steady trickle of 0-25 new subscribers each week, and sometimes gets a large increase. Here are all of the weeks in which there were >25 new subscribers:

AN #1 -> AN #2: 2 -> 141 subscribers (+139), because of the initial announcement.

AN #3 -> AN #4: 148 -> 238 subscribers (+90), probably still because of the initial announcement, though I don't know why it grew so little between #2 and #3.

AN #14 -> AN #15: 328 -> 405 subscribers (+77), don't know why (though I think I did know at the time)

AN #16 -> AN #17: 412 -> 524 subscribers (+112), because of Miles Brundage's tweet on July 23 about his favorite newsletters.

AN #17 -> AN #18: 524 -> 553 subscribers (+29), because of this SSC post on July 30 and the LessWrong curation of AN #13 on Aug 1.

AN #18 -> AN #19: 553 -> 590 subscribers (+37), because of residual effects from the past two weeks.

AN #30 -> AN #31: 653 -> 689 subscribers (+36), because of Rosie Campbell's blog post on Oct 29 about her favorite newsletters.

Over time, the opens and clicks have gone down as a percentage of subscribers, but have gone up in absolute numbers. I would guess that the biggest effect is that the most interested people subscribed early, and so as time goes on the marginal subscriber is less interested and ends up bringing down the percentages. Another effect would be that over time people get less interested in the newsletter, and stop opening/clicking on it, but don't unsubscribe. However, over the last few months, rates have been fairly stable, which suggests this effect is negligible.

On the other hand, during the last few months growth has been organic / word-of-mouth rather than through "publicity" like Miles's tweet and Rosie's blog post, so it's possible that organic growth leads to more interested subscribers who bring up the rates, and this effect approximately cancels the decrease in rates from people getting bored of the newsletter. I could test this with more fine-grained data about individual subscribers but I don't care enough.

So far, I have not been trying to publicize the newsletter beyond the initial announcement. I'm still not sure of the value of a marginal reader obtained via "publicity". The newsletter seems to me to be both technical and insider-y (i.e. it assumes familiarity with basic AI safety arguments), while the marginal reader from "publicity" seems not very likely to be either. That said, I have heard from a few

readers that the newsletter is reasonably easy to follow, so maybe I'm putting too much weight on this concern. I'd love to hear thoughts in the comments.

Composition of subscribers

I don't know who these 889 subscribers are; it's much larger than the size of the field of AI safety. Even if most of the technical safety researchers and strategy/policy researchers have subscribed, that would only get us to 100-200 subscribers. Some guesses on who the remaining people are:

- There are lots of people who are intellectually interested in AI safety but don't work on it full time; maybe a lot of them have subscribed.
- A lot of technical researchers are interested in AI ethics, fairness, bias, explanations and so on. I occasionally cover these topics. In addition, if you're interested in short-term effects of AI, you might be more likely to be interested in the long-term effects as well. (Mostly I'm putting this down because I've met a few people in this category who expressed interest in the newsletter.)
- Non-technical researchers interested in the effects of AI might plausibly find it useful to read the newsletter to get a sense of what AI is capable of and how technical researchers are thinking about safety.

Regardless of the answer, I'm surprised that these people find the newsletter valuable. Most of the time I'm writing to technical safety researchers, and relying on an assumption of shared jargon and underlying intuitions that I don't explain. It's not as bad as it could be, since I try to make my explanations accessible both to people working in traditional AI as well as people at MIRI, but I would have guessed that it was still not easy to understand from the outside. Some hypotheses, only the first of which seems plausible:

- I'm wrong about how difficult it is to understand the newsletter. Perhaps people can understand everything, or maybe they can still get a useful gist from summaries even if they don't understand everything.
- People use it only as a source of interesting papers, and ignore the summaries and opinions (because they are hard to understand).
- Reading the summaries and opinions gives the illusion of understanding even though people don't actually understand what I'm saying.
- People like to feel like a part of an elite group who can understand the technical jargon, and reading the newsletter gives them that feeling. (This would not be a conscious decision on their part.)

I sampled 25 people uniformly at random from the subscribers. Of these, I have met 8 of them, and have heard of 2 more. I would categorize the 25 people in the following rough categories: x-risk community (4), AI researchers sympathetic to x-risk (2), students (3), people interested in AI and x-risk (3), people involved with AI startups (2), researcher with no publicly obvious interest in x-risk (6), and could not be found easily (5). But really the most salient outcome was that for anyone I didn't already know, I found it very hard to figure out why they were subscribed to the newsletter.

Impact of the newsletter

In which I try and fail to figure out whether the benefits outweigh the costs.

Benefits

Here are the main sources of value from the newsletter that I see:

- Causing technical researchers to know more about other areas of the field besides their own subfield.
- Field building, by giving new entrants into AI safety a way to build up their knowledge without requiring mentorship.
- Improving the reputation of the field of AI safety (especially among the wider AI research community), by demonstrating a level of discourse above the norm, particularly in conjunction with good writing about current AI topics. There's a mixture of reasoning about current AI and speculative future predictions that clearly demonstrates that I'm not some random outsider critiquing AI researchers.
- Creating a strong reputation for myself and CHAI, such that people will have justified reason to listen to CHAI and/or me in the future.
- Providing some sort of value to the subscribers who are not in long-term AI safety or AI strategy/policy.

When I started the newsletter, I was aiming primarily for the first one, by telling researchers what they should be reading. I continue to optimize mainly for that, though now I often try to provide enough information that researchers don't have to read the original paper/post. I knew about the second source of value, but didn't think it would be very large; I'm now more uncertain about how important it is. The reputational effects were more unexpected, since I didn't think the newsletter would become as large as it currently is. I don't know much about the last source of value and am basically ignoring it (i.e. pretending it is zero) in the rest of the analysis.

I'm actually quite uncertain about how *much* value comes from each of these subpoints, mainly because there's a striking lack of comments or feedback on the newsletter. Excluding one person at CHAI who I talk to frequently, I get a comment on the content of the newsletter maybe once every 3-4 weeks. I can understand that people who get it as an email newsletter may not see an obvious way to comment (replying to a newsletter email is an unusual thing to do), but the newsletter is crossposted to LessWrong, the Alignment Forum, and Twitter. Why aren't there comments there?

One possibility is that people treat the newsletter as a curation of interesting papers and posts, in which case there isn't much need to comment. However, I'm fairly confident that many readers also find value in the summaries and opinions. You could instead interpret this as evidence that the things I'm saying are reasonable -- after all, if I was wrong on the Internet, surely someone would let me know. On the other hand, if I'm only saying things that people already believe, am I actually accomplishing anything? It's hard to say.

I think the most likely story is that I say things that people didn't know but agree with once I say them -- but I share Raemon's intuition that people aren't really learning much if that's the case. (The rest of that post has many more thoughts on comments that apply to the newsletter.)

Overall it still feels like in expectation most of the value comes from widening the set of fields that any individual technical researcher is following, but it seems entirely possible that the newsletter does not do that at all and as a result only has reputational benefits. (I am fairly confident that the reputational benefits are positive

and non-zero.) I'd really like to get more clarity on this, so if you read the newsletter, please take the survey!

Costs

The main cost of the newsletter is the opportunity cost of our time. Each newsletter takes about 15 hours of my time. The newsletter has gotten more detailed over time, but this isn't reflected in the total hours I put in because it has been approximately offset by new content writers (Richard Ngo and Dan Hendrycks) who took some of the burden of summarizing off of me. Currently I'd estimate that the newsletter takes 15-20 hours in total (with 2-5 hours from Richard and Dan). This can be broken down into time I would have spent reading and summarizing papers anyway, and time that I spent only because the newsletter exists, which we could call "extra hours". Initially, I wanted to read and summarize a lot of papers for my own benefit, so the newsletter took about 4-5 extra hours per week. Now, I'm less inclined to read a ton of papers, and it take 8-10 extra hours per week.

This means in aggregate I've spent 700-800 hours on the newsletter, of which about 300-400 were hours that I wouldn't have spent otherwise. Even only counting the 300-400 hours, this is comparable to the time I spent on state of the world and learning biases projects *together*, including all of the time spent on paper writing, blog posts, and talks in addition to the research itself.

In addition to time costs, the newsletter could do harm. While there are many ways this *could* happen, the only one that feels sufficiently important to consider is the risk of causing information cascades. Since nearly everyone in the field is reading the newsletter, we may all end up with some belief B just because it was in a newsletter. We might then have way too much confidence in B since everyone else also believes B.

Overall I'm not too worried. There's so much content in the newsletter that I seriously doubt a single idea could spread widely as a result of the newsletter -- inevitably some people won't remember that particular idea. So we only need to worry about "big" ideas that are repeated often in the newsletter. The most salient example of that would be my general opposition to the Bostrom/Yudkowsky paradigm of AI safety, but it still seems quite prevalent amongst researchers. In addition I'd be really surprised if existing researchers were convinced of a "big" idea or paradigm solely because other researchers believed it (though they might put undue weight on it).

Is the newsletter worth it?

If the only benefit of the newsletter were the reputational effects, it would not be worth my time (even ignoring Richard and Dan's time). However, I get enough thanks from people in the field that the newsletter must be providing value to them, even though I don't have a great model of what the value is. My current best guess is that there is a lot of value, which makes the newsletter worth the cost, but I think there is a non-negligible chance that this would be reversed if I had a good model of what value everyone was getting from it.

Going forward

In which I figure out what about the newsletter should change in the future.

Structure of the newsletter

So far I've only talked about whether the newsletter is worthwhile as a whole. But of course we can also analyze individual aspects of the newsletter and figure out how important they are.

Opinions are probably the key feature of the newsletter. Many papers and blog posts are aimed more at appearing impressive rather than conveying facts. Even the ones that are truth seeking are subject to publication bias: they are written by people who think that the ideas within are important, and so will be biased towards positivity. As a result, an opinion from a researcher who *didn't* do the work can help contextualize the results that makes it easier for less involved readers to figure out the importance of the ideas. (As a corollary, I worry about the lack of a fresh perspective on posts that I write, but don't see an obvious easy solution to that problem.) I think this also contributes to the success of Import AI and ChinAI, which are also quite heavy on opinions.

I think the summaries are also quite important. I aim for the longer summaries to be sufficiently informative that you don't have to read the blog post / paper unless you want to do a deep dive and really understand the results. For papers, I often roughly aim for it to be more useful to read my summary than to read the abstract, intro, and conclusion of the paper. In the world where the newsletter didn't have summaries, I think researchers would not keep up as much with the state of the field.

Overall, I think I'm pretty happy with the current structure of the newsletter, and don't currently intend to change it. But if I get more clarity on what value the newsletter provides to researchers, I wouldn't be surprised if I would change the structure as a result.

Scaling up

In the year that I've been writing the newsletter, the amount of writing that I want to cover has gone up quite a lot, especially with the launch of the Alignment Forum. I expect this will continue, and I won't be able to keep up.

By default, I would cover less and less of it. However, it would be nice for the spreadsheet to be a somewhat comprehensive database of the AI safety literature. This is not what we currently have, because I often don't cover good Agent Foundations work because it's hard for me to understand and I don't have pre-2018 content, but it is pretty good for the subfields of AI safety that I'm most knowledgeable about.

There has been some outsourcing of work as Richard Ngo and Dan Hendrycks have joined, but it still does not seem sustainable to continue this long-term, due to coordination challenges and challenges with maintaining quality. That said, it's not impossible that this could work:

- Perhaps I could pay people to do this summarization, with the hope that this would help me find people who could put in more time. This would allow more work to get done while keeping the team small (which keeps coordination costs and quality maintenance costs small).

- I could create a system that allows random people to easily contribute summaries of papers and posts they have read, while writing the opinions myself. It may be easier to vet and fix summaries than to write them myself.
- I could invest in developing good guides for new summarizers, in order to decrease the cost of onboarding and ongoing coordination.

That said, in all of these cases, it feels better to instead just summarize a smaller fraction of all the work, especially since the newsletter is already long enough that people probably don't read all of it, while still adding links to papers that I haven't read to the spreadsheet. The main value of summarizing everything is having a more comprehensive spreadsheet, but I don't think this is sufficiently valuable to warrant the approaches above. That said, I could imagine that this conclusion being overturned by having a better model of how the newsletter adds value for technical safety researchers.

Sourcing

So far, I have found papers and articles from newsletters, blogs, Arxiv Sanity and Twitter. However, Twitter has become worse over time, possibly because it has learned to show me non-academic stuff that is more attention-grabbing or controversial, despite me trying not to click on those sorts of things. Arxiv Sanity was my main source for academic work, but recently it's been getting worse, and is basically not working any more, and I'm not sure why. So I'm now trying to figure out a new way to find relevant literature -- does anyone have suggestions?

If I continue to have trouble, I might summarize random academic papers I'm interested in instead of the ones that have come out very recently.

Appearance

It's rather annoying that the newsletter is a giant wall of text; it's probably not fun to read as a result. In addition to the categories, which were partly meant to give structure to the wall of text, I've been trying to break things into more paragraphs, but really it needs something much more drastic. However, I also don't want it to be even more work to get a newsletter out.

So, if anyone wants to volunteer to make the newsletter visually nicer that would be appreciated, but it shouldn't cost me too much more time (maybe half an hour a week, if it was *significantly* nicer). One easy possibility would be to include an image at the beginning of the newsletter -- any suggestions for what should go there?

Future of the newsletter

Given the uncertainty of the value of the newsletter, it's not inconceivable that I decide to stop writing it in the future, or scale back significantly. That said, I think there is value in stability. It is generally bad for a project to have "fits and starts" where its quality varies with the motivation of the person running them, or for the project to potentially be cancelled solely based on how valuable the creator thinks it is. (I'm aware I haven't argued for this; feel free to ask me about it if it seems wrong.)

Due to this and related reasons, when I started the newsletter, I had an internal commitment to continue writing it for at least six months, as long as most other

people thought it was still valuable. Obviously, if everyone agreed that the newsletter was not useful or actively harmful, then I'd stop writing it: this is more to deal with the case where I no longer think the newsletter is useful, even though other people think it is useful.

Now I'm treating it as an ongoing three-month commitment: that is, I am always committing to continue writing the newsletter for at least three months as long as most other people think it is valuable. At any point I can decide to stop the ongoing commitment (presumably when I think it is no longer worth my time to write it); there would then be three months where I would continue to write the newsletter for stability, and figure out what would happen with the newsletter after the three months.

Feedback I'd like

There are a bunch of questions I have, that I'd love to get opinions on either anonymously in the 3-minute survey (which you should fill out!) or in the comments. (Comments preferred because then other people can build off of them.) I've listed the questions roughly in order of importance:

- **What is the value of the newsletter for you?**
- What is the value of the newsletter for other people?
- How should I deal with the growing amount of AI safety research?
- What can I do to get more feedback on the newsletter on an ongoing basis (rather than having to survey people at fixed times)?
- Am I underestimating the risk of causing information cascades? Regardless, how can I mitigate this risk?
- How can I make the newsletter more visually appealing / less of a wall of text, without expending too much weekly effort?
- Should I publicize the newsletter on Twitter? How valuable is the marginal reader?
- Should I publicize the newsletter to AI researchers? How valuable is the marginal reader?
- How can I find good papers out of academia now that Arxiv Sanity isn't working as well as it used to?

Appendix: Alignment Newsletter FAQ

All of these are in the appendix because I don't particularly care if people read it or not. It's not very relevant to any of the content in the main post. It *is* relevant to anyone who might want to start their own newsletter, or their own project more generally.

What's the history of the Alignment Newsletter?

During one of the CHAI seminars, someone suggested that we each take turns finding and collecting new research papers and sending them out to each other. I already had a system in place doing exactly this, so I volunteered to do this myself (rather than taking turns). I also figured that to save even more CHAI-researcher-time, it would make sense to give a quick summary and then tell people under what circumstances they should read the paper. (I was already summarizing papers for my own notes.)

This pretty quickly proved to be valuable, and I thought about making it public for even more time savings. However, it still seemed pretty nascent and in flux, so I continued iterating on it within CHAI, while thinking about how it could be made to be public-facing. (See also the “Things done right” section.) After a little under two months of writing the newsletter within CHAI, I made it public. At that time, the goal was to provide a list of relevant readings for technical AI safety researchers that had been published each week; and help them decide whether or not they should read them.

Over time, my summaries and opinions became longer and more detailed. I don’t know exactly why this happened. Regardless, at some point I started aiming for some of my summaries to be detailed enough that researchers could just read the summary and not read the paper/post itself.

In September, Richard Ngo volunteered to contribute summaries to the newsletter on a variety of topics, and Dan Hendrycks joined soon after focusing on robustness and uncertainty.

Why do you never have strong negative opinions?

One of the design decisions made at the beginning of the newsletter was to avoid strong critiques of any particular piece of research. This was for a few reasons:

- As a general rule, any criticism I have of a paper is often too strong or based on a misunderstanding. If I have a negative impression of a paper or research agenda, I would predict that with ~90% probability after I talk to the author(s) my opinion of the work will have improved. I don’t think this is particular to me -- this should be expected of any summarizer since the authors have much more intuition about why their particular approach will be useful, beyond what is written in the blog post or paper.
- The newsletter probably shapes the views of a significant fraction of people thinking about AI safety, and so leads to a risk of information cascades. Mitigating this means giving space to views that I disagree with, summarizing them as best I can, and not attacking what will inevitably be a strawman of their view.
- Regardless of the accuracy of the criticism, I would like to avoid alienating people.

Of course, this decision has downsides as well:

- Since I’m not accurately saying everything I believe, it becomes more likely that I accidentally say false things, convey wrong impressions, or otherwise make it harder to get to the truth.
- Disagreements are one of the main ways in which intellectual progress is made. They help identify points of confusion, and allow people to merge their models in order to get something (hopefully) better.

While the first downside seems like a real cost, the second downside is about inhibiting *intellectual progress* in AI safety research. I think this is okay: intellectual progress does not need to happen in the newsletter. In most of these cases I express stronger disagreements in channels more conducive to intellectual progress (e.g. the Alignment Forum, emails/messages, talking in person, the version of the newsletter internal to CHAI).

Another probable effect of avoiding negativity is reduced readership, since it is likely much more interesting to read a newsletter with active disagreements and arguments than one that dryly summarizes a research paper. I don't yet know whether this is a pro or a con (even ignoring other effects of negativity).

Mistakes

I don't know of very many mistakes, even in hindsight. I think this is primarily because I don't get feedback on the newsletter, not because everything has gone perfectly. It seems quite likely that there are still things that are mistakes; but I don't know it yet because I don't have the data to tell.

Analyzing other newsletters. The one thing that I wish I had done was to analyze other newsletters like Import AI in more detail before starting this one. I think it's plausible that I could have realized the value of opinions and more detailed summaries right at the beginning, rather than evolving in that direction over a couple of months.

Delays. I did fall over a week behind on the newsletter over the last month or two. While this is bad, I wouldn't really call it a Mistake: I don't think of the newsletter as a weekly commitment or obligation. I very much value the flexibility to allocate time to whatever seems most pressing; if the newsletter was more of a commitment (such that falling behind is a Mistake), I think I would have to be much more careful about what I agree to do, and this would prevent me from doing other important things. Instead, my approach is to have the newsletter as a fairly important goal that I try to schedule enough time for, but if I find myself running out of time and have to cut something, it's not a tragedy if it means the newsletter is delayed. That's essentially what happened over the last month or two.

Things done right

I spent a decent amount of time thinking about the design of the newsletter before implementing it, and I think this was in hindsight a very good idea. Here I list a few things that worked out well.

A polished product. I was particularly conscious of the fact that at launch the newsletter would be using up the limited common resource of "people's willingness to try out new things". Both in order to make sure people stuck with the project, and in order to not use up the common resource unnecessarily, I wanted to be fairly confident that this would be a good product before launching. As a result, I iterated for a little under two months within CHAI, in order to figure out product-market fit. You can see the evolution over time -- this is the first internal newsletter, whereas this is the first public newsletter. (They're all available here.)

- By the fourth internal newsletter, I realized that I couldn't actually summarize all the links I found, so I switched to a version where some links would be sent without summaries.
- Categorization seemed important, so I did more of it.

This is not to say that the newsletter has been static since launch; it has changed significantly. Most notably, while originally I was aiming to give people enough information to decide whether or not to read the paper/post, I now sometimes aim for including enough detail that people don't need to read the paper/post. But the point is

that a lot of the early improvements happened within CHAI without consuming the common resource.

I'm not sure to what extent this is different from standard startup advice of iterating quickly and testing product-market fit: it depends on whether it counts as testing for product-market fit to trial the newsletter within CHAI. To the extent that there is a difference, it's mainly that I'm arguing for more planning, especially before consuming common resources (whereas with startups, the fierce competition means that you do not worry about consuming common resources).

Considered stability and commitment. As I mentioned above, I had an internal commitment to continue writing the newsletter for at least six months, as long as other people thought it was valuable. In addition to the value of stability, I viewed this as part of cooperatively using the common resource of people's willingness to try things. If you're going to use the resource and fail, ideally you would have learned that it is actually infeasible to succeed in that domain, as opposed to e.g. lack of motivation on the author's part.

Here's another way to see this. I think it would have been a lot harder for the newsletter to be successful if there had been 2-5 attempts to create a newsletter in the past that had then fizzled out, because people would expect newsletters to fail and wouldn't subscribe. My initial commitment helps prevent me from being one of those failures for "bad" reasons (e.g. me losing motivation) while still allowing me to fail for "good" reasons (e.g. no one actually wants to read a newsletter about AI alignment).

I can't point to any actually good outcomes that resulted from this policy; nonetheless I think it was a good thing to have done.

Investing in *flexible* automated systems. I had created the private version of the spreadsheet before the first public newsletter, in order to have a database of readings for myself (replacing my previous Google Doc database), and I wrote a script to generate the email from this database. While lots of ink has been spilled on the value of automation, it doesn't usually emphasize *flexibility*. By not using a technology meant for one specific purpose, I was able to do a few things that I wouldn't expect to be able to do with a more specialized version:

- Create consistency checks. For example, throwing an error when there's an opinion but no summary, or when the name of the summarizer is not "Richard", "Dan H" or "" (indicating me).
- Creating a private and public version of the newsletter. (Any strong critiques go into the private version, which is internal to CHAI, and are removed from the public version.)

But really, the key value of flexibility is that it allows you to adapt to circumstances that you had never even considered when creating the system:

- When Richard Ngo joined, I added a "Summarizer" column to the sheet, changed a few lines of code, and was done. (Note how I needed flexibility over both the data format and the analysis code.)
- I've found myself linking to a bunch of previous newsletter entries and having to copy a lot of links. Recently I added a new tag that I can use in summaries and opinions that automatically extracts and links the entry I'm referring to. (I'm a bit embarrassed at how long it took me to realize that this was a thing I could

do; I could have saved a lot more tedious work if I had realized it was a possibility the first time I got annoyed at this process.)

Thought about potential negative effects. I'm pretty sure I thought of most of the points about negativity (listed above) before publicizing the newsletter. This is discussed a lot; I don't think I have anything significant to add.

This section seems to indicate that I thought of things initially and they were all important -- this is almost certainly not the case. I'm sure I'm rationalizing some of these with hindsight and didn't actually think of all the benefits then, and I also probably thought of other considerations that didn't end up being important that I've now forgotten.

Prompts for eliciting blind spots/bucket errors/bugs

This post is to make publicly available a few prompts/questions I came up with aiming to uncover blind spots around identity/self-concepts.

- Select a trait X that you believe you have, and where you like that you have it (e.g. rational, kind, patient...)
- Try to imagine a character that is a caricature of someone with trait X. Or another way to think about this: The way Spock is a Straw Man version of a rational character, what would a Straw Man version of a character with trait X look like? (referred to in the following as X-Spock)
    - What are blind spots an X-Spock is likely to have?
    - In what sorts of situations is an X-Spock especially likely to fail?
    - What would an X-Spock have a lot of trouble admitting to? (e.g. someone who considers themselves courageous may be unable to admit they are afraid)
    - What are traits that seem like opposites of X?
        - Could the opposite traits actually be beneficial?
        - Is what seems like an opposite trait in actuality orthogonal? (e.g. rational and emotional)

Conditional revealed preference

This is a linkpost for https://unstableontology.com/2019/04/04/conditional-revealed-preference/

In the linked post, I discuss revealed preference analysis and argue that claims about someone's actual values should predict not only what they do in their current situation, but what they would do in substantially different situations, given e.g. different information and different expectations for how others behave.

(LessWrongers may ask "what does this add over CEV?" CEV is one possible hypothetical where people are more informed, but is impossible to compute and also not predictive of what people do in actual situations of having somewhat more information while still being under cognitive limitations. Thus, CEV "analysis", unlike conditional/counterfactual analysis, ends up being largely spurious.)

Liar Paradox Revisited

A well-known brainteaser asks about the truth of the statement "this statement is false". My previous article on this topic, outlined common approaches to this problem and then argued that we should conceive of two distinct kinds of truth:

- Statements about the world, where as per Tarski there is a natural interpretation: "Snow is white" is true iff and only iff "Snow is white"
- Logical/Mathematical statements, where the notion truth is constructed to give us a convenient way of talking about our rules of inference within a particular system that normally excludes self-referential statements

I should add that some statements can only be defined using a combined notion of truth, ie. "The car is red and 1+1=2".

My point was that if we chose to extent logical/mathematical statements outside of their usual bounds, we shouldn't be surprised that it breaks down and that if we choose to patch it, there will be multiple possible ways of achieving this.

Patching with INFINITE-LOOP

So let's consider how we might attempt to patch it. Suppose we follow the Formalists (ht CousinIt) and insist that "true" or "false" or only applied to sentences that can be evaluated by running a finite computation process. Let's add a third possible "truth" value: INFINITE-LOOP.

Consider the following sentence:

```
            The truth value of this sentence is not INFINITE-LOOP
```

This seems to be a contradiction because the sentence is infinitely recursive, but at the same time denies this.

In order to understand what is happening, we need to make our algorithm for assigning truth values more explicit:

```
            if expansion terminates:
    Resolve truth value by expanding expressions with references to other
expressions
else:
    Assign INFINITE-LOOP if expansion fails to terminate
```

What we see here is that if the sentence is not able to be expanded without ending up in an infinite loop, it is assigned the truth value INFINITE-LOOP without any regard to what the sentence asserts. So there isn't actually an inconsistency, at most, this system for assigning truth values just isn't behaving how we'd want.

In fact consider the following:

```
            A: This sentence is false
B: Sentence A has a truth value of INFINITE-LOOP
```

According to the above algorithm, assigning INFINITE-LOOP to B is correct, when it seems like it should be FALSE. Further, this system assigns INFINITE-LOOP to:

```
            1+1=2 or this sentence is false
```

when perhaps it'd be better to assign it a value of TRUE.

Patching with an oracle

Being able to talk about whether or not sentences end up in an infinite loop seems useful. So we can imagine that we have a proof oracle that can determine whether sentence will end up in a loop or not.

```
            for reference in sentence:
    if oracle returns INFINITE-LOOP:
         Evaluate the clause given the value INFINITE-LOOP as the truth value of the
reference
    else:
        Expand normally
```

However, our oracle still doesn't demystify:

```
            The truth value of this sentence is not INFINITE-LOOP
```

As our algorithm would replace the first clause with INFINITE-LOOP and hence evaluate

```
            INFINITE-LOOP is not INFINITE-LOOP
```

to FALSE. But then:

```
            FALSE is not INFINITE-LOOP
```

so we would expect it to also be TRUE.

So perhaps we should define our oracle to only work with sentences that don't contain references to INFINITE-LOOPS. Consider the following situation:

```
            A: This sentence is false
B: Sentence A has a truth-value of INFINITE-LOOP
C: Sentence B is true
D: Sentence C is true
```

B would be TRUE (even though it refers to INFINITE-LOOP, the oracle only has to work with the reference "Sentence A"). However, C would be undefined.

We could fix this by allowing the oracle to return TERMINATES for sentences that can be evaluated after one level of expansion with our initial definition of an oracle. We can then allow sentence D to be true by allowing the oracle to return TERMINATES for

any sentence that can be evaluated after two levels of expansion and we can recursively extend this definition until infinity.

This also resolves cases like:

```
1+1=2 or this sentence is false
```

The second clause evaluates to INFINITE-LOOP and since this is a truth value, rather than actually infinitely looping, (TRUE OR INFINITE-LOOP) should give true.

Patching with ORACLE-LOOP

We still haven't figured out how to handle cases like:

```
The truth value of this sentence is not INFINITE-LOOP
```

I would suggest that we might want to repeat our first move and say that the truth value is ORACLE-LOOP whenever an oracle fails to resolve it (even if we expand it an infinite number of times, we still end up with a sentence containing INFINITE-LOOP). We can then stack meta-levels and further metalevels on top of this.

Fixed points

We could also define another notion of truth whether a statement is true when there is a single fixed point. This would result in statements like:

```
This sentence is true
```

Being set to true instead of INFINITE-LOOP.

In any case, the way that we extend the concept of truth to apply to these degenerate cases is purely up to what we find convenient.

Final Note

I actually think there's a similarity here to the Least Interesting Number Paradox. This paradox seems paradoxical only when we fail to define interesting precisely. Once we start to do that, we realise that Interesting Meta-Level 0 and Interesting Meta-Level 1 are different. Similarly, there are different levels of truth depending on how far we go in trying to resolve a question.

Against Street Epistemology

According to https://streetepistemology.com/publications/street_epistemology_the_basics , street epistemology is a "conversational technique" which is intended to be "a more productive and positive alternative to debates and arguments." Street epistemologists assume a role similar to Socrates in Plato's dialogues, asking questions of his interlocutor to try to create a realisation of ignorance in them. The goal of street epistemology is to find incoherences in people's beliefs, and to convince them of the value of "scepticism."

A street epistemologist tries to remain calm and pleasant throughout the entire interaction, and to build rapport at the beginning in order to make their interlocutor comfortable with the exchange. After introductions and rapport are established,they can ask their interlocutor to identify a belief and give an approximate level of confidence in it (on a scale of 1 to 10). The early stages of the conversation, after identifying the belief, are devoted to making the belief clear and precise so that there is as little ambiguity as possible, and less wiggle room if and when incoherences are found. Terms are defined, clarifying questions are asked and answered. To confirm that the belief is understood, before trying to undermine it, the street epistemologist will try to give a paraphrase of the view that his interlocutor finds charitable and acceptable.

Having pinpointed what the claim is, the street epistemologist then asks which methods the interlocutor used to arrive at their confidence level in this belief. This is the very first question in what might be called the cross-examination stage, and it reveals what sort of incoherence is being sought in these conversations. Street epistemology is all about finding poorly articulated or unarticulated spots in people's epistemological views. One the interlocutor has given a few answers and it comes time to dive into them, the website recommends focusing only on "one or two" of the methods listed, ideally those that are identified as most important. The idea here is to ask many questions about the methods and sources of evidence, hoping that at some point there will be a stumper that causes the interlocutor to notice mistakes or incoherence in their thought. Whether the street epistemologist succeeds is identified partly by observing their body language during the interaction: the website mentions, for example, that looking up in thought to avoid eye contact might be a sign.

Crucially, during this whole process the street epistemologist is told to avoid talking about the content of the beliefs or the particulars of the evidence. Again, the point is to find holes in the interlocutor's epistemology, not per se with the belief itself. The conversation is a step removed from the belief, instead focusing on examining the epistemology on which the belief is supposedly based. For that reason, street epistemologists are also instructed to avoid presenting additional evidence or advancing arguments against the interlocutor's position. To do any of this is described on the website as getting "pulled in the weeds" and "sidetracked." The questions and of the conversation should only be about the one or two methods chosen earlier.

That's enough of a sketch that I can dive into my thoughts. Street epistemology strikes me as misguided in a couple of ways. First, because it treats scepticism as a positive position or worldivew, abstracted from the intentional content , rather than an attitude toward particular beliefs or proposals. ("Sceptical about what?" "Oh, just sceptical in general.") Second, because it assumes that people's ability to articulate

and defend epistemological methods has any special bearing on the warrant for their belief.

On the first, we've already seen that people who practice street epistemology have the explicit aim of spreading scepticism, and that in doing so they consider dealing with the particular content of the beliefs "getting stuck in the weeds." The mode of conversation, the Socratic chain of questions and answers, is abused in order to avoid having to deal with the particular evidence or content. Usually, someone is sceptical t*oward* some belief f*or* some reason. Those reasons will normally have to do with the content of the belief, the associated evidence or arguments, and so forth. Yet street epistemology suggests, by what it chooses to ask about and through the premises of the questions, that people should be sceptical that they can know anything. It suggests that the methods they use are so faulty that knowing the particulars of the belief is extraneous information.

It is only after the series of questions successfully undermine the interlocutor's sense of their epistemic foundations, not just for the belief in question, but in general, that the conversation ever goes back to the particulars and context of the initial belief. And then, it is only to point out that the interlocutor's responses to the questions imply that it is unwarranted. The problem with this is that in any earnest form of enquiry, the relevant context of the particular judgements has to be considered from the very beginning. A claim, even if successful, that one of several methods was faulty or unreliable does not itself touch on the veracity of the belief.

The point here is analogous to one often made about logical fallacies. While it may be true that affirming the consequent is invalid, it does not follow that any instance of affirming the consequent is unsound. In order to determine that, we would have to consider the content of the premises and the balance of the evidence for or against them. Indeed, going into those details is also necessary to determine whether the particular method in this case was faulty. Reconstructing an argument presented in natural language into a symbolic form requires grappling with the meaning of the claims. Logical structure can be discerned only by knowing the meaning and content.

On the second issue, it may be the case that the interlocutor has never made their methods of knowing sufficiently explicit to articulate them accurately in a conversation, if they are even aware of vaguely what those methods are. It depends on how experienced they are at thinking about the particular topic at hand. Because of this, they might mischaracterise the relevant methods in order to provide a enough detail to satisfy the norms of the conversation. Asking a relatively intellectually inexperienced person (on the topic at hand) to elaborate on their methods is a bit like trying to ask a novice at the gym why he leaned forward when going down into a squat, and whether that was good or bad. There are definite reasons that any more experienced powerlifter will perceive right away, but the novice won't know where to look in his experience to find the answer (although it is there, in his experience).

Rather than asking them outright to give explicit reasons for a belief for the purpose of questioning them, it would be much better to have them give several minutes of uninterrupted, free-association thoughts about the subject, and then try to tease out their reasoning from there. This closely resembles the best practices in witness interviews developed by forensic psychologists. It's the best way to ensure that you are not relying on a potentially non-existent explicit understanding in your questions, and to avoid biasing their statement about the beliefs and evidence with the wording and direction of your questions.

This gets at a broader point which people tend not to consider, which is that most forms of theoretical knowledge are more like the knowledge of how to play the piano than they are like knowledge of what you ate for breakfast. They are driven by craft, which is passed down by both written and oral traditions through training and schooling.

Training in biology, for example, involves spending time having the best techniques, mnemonics, and interpretations of the relevant material or activity explained to you by experts who have already mastered them to some degree. Learning and remembering facts is part of it, but there is also skill and experienced judgement in how a biologist designs an experiment, in how they set up the laboratory equipment to conduct it, and in how they interpret and organise the results for publication. There needs to be a sense, at every step, of what's appropriate, what's relevant, what's fruitful, and so forth. This also applies to our ability to use the five senses in extremely normal ways. Something as quintessential as noticing what's in your visual field is an activity which you can spend your entire lifetime honing and teaching, as visual artists do.

Street epistemology tries very hard to be neutral in several ways: with respect to evidence, arguments, the tone and tenor of the conversation, and even the particular methods that are called into question. One motivation behind this is that the conversation is not about the street epistemologists' views, whether epistemic or in terms of the content of the belief in question. The conversation is supposed to be exclusively about the interlocutor's epistemic methods relative to one of their views, methods which are explored in such a way that their internal contradictions reveal themselves over the course of impartial questions.

What I've tried to argue, though, is that street street epistemology smuggles in many assumptions about epistemology. It is hardly neutral about it in the way it assumes and requires. By avoiding the context and content of the beliefs, street epistemology adopts a view in which knowledge is a nexus of propositions connected by chains of reasoning. These chains are formed by "epistemic methods" (analogous with proposed rules of inference) which may or may not be reliable (analogous with valid). Someone is warranted to believe a proposition only if they can show that they arrived at it by a reliable method. Because of this, people of varying levels of skill and practice in the relevant activities are held to the same epistemic expectations. They are all equally expected to articulate their beliefs in such a way that they can identify, not only the evidence for them, but also some methods used to obtain and assess the evidence. They must also be able to defend these methods as generally reliable, when abstracted from the context of the particular belief. Even if this standard were good, which I hope I have at least called into question, this would be unreasonable. People with little to no ability to engage in epistemological discourse may still be warranted to have beliefs.

Street epistemology tries to characterise the conversation as conceptually neutral. Yet its take on the Socratic method bakes in so many assumptions about epistemology that it is very easy to disagree, in at least four ways: a) with the implicit epistemology, b) with the characterisation of the discourse implicit in the premises of the questions, c) with the aim and structure of the line of questioning, and d) with the content of the questions, including the content excluded by them. I provided examples of disagreement along all these lines in the discussion of craft above. To address any of these problems, though, requires stepping outside of the framework of street epistemology, because it requires that the street epistemologist, the questioner,

express positive views. This ends the Socratic mode of a series of questions and answers, and gets "stuck in the weeds" of discussing the content of beliefs.

In order to seek, in good faith, to help improve an interlocutor's beliefs, I would suggest doing exactly what street epistemologists are advised against in the online guide. Go to the particulars of the belief and the relevant context. Engage directly in argument. Look at varying examples of evidence. Do not be afraid of frustration or heated exchanges, since often that is an important sign for both people that they have reached a crucial problem in the discussion that needs to be resolved. Teach relevant skills if the other person lacks them. Perhaps most important, expect the conversation to last off and on for a long time in the context of acquaintance or friendship, rather than as a one-off exchange with a stranger.

Machine Pastoralism

This idea has occurred to me before, but in the interim I dismissed it and then forgot. Since it is back again more-or-less unprompted, I am writing it down.

We usually talk about animals and their intelligence as a way to interrogate intelligence in general, or as a model for possible other minds. It occurred to me our *relationship* with animals is therefore a model for our relationship with other forms of intelligence.

In the mode of Prediction Machines, it is straightforward to consider: prediction engines in lieu of dogs to track and give warning; teaching/learning systems for exploring the map in lieu of horses; analysis engines to provide our solutions instead of cattle or sheep to provide our sustenance. The idea here is just to map animals-as-capital to the information economy, according to what they do for us.

Alongside what they do for us is the question of how we manage them. The Software 2.0 lens of adjusting weights to search program space reads closer to animal husbandry than building a new beast from the ground up with gears each time, to me. It allows for a notion of lineage, and we can envision using groups of machines with subtle variations, or entirely different machines in combination.

This analogy also feels like it does a reasonable job of priming the intuition about where dangerous thresholds might lie. How smart is smart enough to be dangerous for one AI? Tiger-ish? We can also think about relative intelligence: the primates with better tool ability and more powerful communication were able to establish patronage and then total domestication over packs of dogs and herds of horses, cattle, and sheep. How big is that gap exactly, and what does that imply about the threshold for doing the same to humans? Historically we are perfectly capable of doing it to ourselves, so it seems like the threshold might actually be lower than us.

The Stack Overflow of Factored Cognition

Abstract

Factored cognition is a possible basis for building aligned AI. Currently Ought runs small-scale experiments with it. In this article I sketch some benefits of building a system for doing large-scale experiments and generating large amounts of data for ML training. Then I estimate roughly how long it would take to build such a system. I'm not confident of this exploration being useful at all. But at least I wrote it down.

Benefits

If you want to know what factored cognition is, see here.

Ought does small-scale experiments with factored cognition (cf. Ought's Progress Update Winter 2018). I thought: wouldn't it be nice to do these experiments at much larger scale? With enough users that one root question could be answered within three hours any time and day of the week.

Benefits:

- The feedback loop would be much tighter than with the weekly or bi-weekly experiments that Ought runs now. A tight feedback loop is great in many ways. For example, it would allow a researcher to test more hypotheses more often, more quickly and more cheaply. This in turn helps her to generate more hypotheses overall.

  Note that I might be misunderstanding the goals and constraints of Ought's experiments. In that case this benefit might be irrelevant.

- It would generate a lot of data. These could be used as training data when we want to train an ML system to do factored cognition.

Quantifying these benefits is possible, but would take some weeks of modelling and talking with people. So far I'm not confident enough of the whole idea to make the effort.

Feasibility

We would need three things for a large-scale factored cognition system to work: the system itself, enough users and useful behaviour of these users. I'll use Stack Overflow as a basis for my estimates and call large-scale factored cognition ‘Fact Overflow’.

Building Stack Overflow took five months from start of development 🈂 to public beta 🈂. Then they spent a lot of time tweaking the system to make it more attractive and

maintain quality. So I'd say building Fact Overflow would take five to fifteen months with a team of two to five people.

For calculating how many users would be required, I used the following estimates (90 % confidence interval, uniformly distributed):

variable	– 5 %	– 95 %	– explanation
n_w	15	300	– average number of workspaces per tree
n_a	1	5	– average number of actions per workspace
x_c	0.1	0.7	– decontamination factor
x_a	0.1	0.7	– share of active users among all users
f_a/d	1	10	– average frequency of actions per active user

(I had to insert dashes to make the table look neat.)

x_c is the share of workspaces in a tree that one user can work on without being contaminated, ie. without getting clues about the context of some workspaces.

The estimates are sloppy and probably overconfident. If people show interest in this topic, I will make them tighter and better calibrated.

Now if we want a tree of workspaces to be finished within t_f, we need n_u^* users, where:

$$n_u^* = \frac{n_w \cdot n_a}{x_c \cdot x_a \cdot f_a \cdot t_f}$$

A Guesstimate model based on this formula tells me that for $t_f = 3\,\text{h}$ we need between 600 and 36 k users. Note that Guesstimate runs only 5000 samples, so the numbers jump around with each page reload. Note also that the actual time to finish a tree might be longer, depending on how long users take for each action and how many sub-questions have to be worked on in sequence.

How long would it take to accumulate these numbers of users? For this I use the number of sign-ups to Stack Exchange (of which Stack Overflow is the largest part). Let me assume that between 75 % and 98 % of people who sign up actually become users. That means between 700 and 42 k sign-ups are required. This is also in Guesstimate. What I can't include in the Guesstimate simulation is the difference between the growth rates of Stack Overflow and Fact Overflow. Assume that it takes Fact Overflow twice as long as Stack Overflow to reach a certain number of sign-ups. Then it would take one month to reach 700 sign-ups and twenty-two months to reach 42 k sign-ups.

Of course, the system would have to be useful and fun enough to retain that many users. As with Stack Overflow, the software and the community have to encourage and ensure that the users behave in a way that makes factored cognition work.

Conclusion

It would be useful to be able to experiment with factored cognition at a large scale. I can't quantify the usefulness quickly, but I did quantify very roughly what it would take: five to fifteen months of development effort with a small team plus one to twenty-two months of accumulating users.

Comment prompts

- What do you think I'm misunderstanding?
- Do you think my exploration of large-scale factored cognition is a waste of time? If so, why?
- Do you think one could build a platform attractive enough to that many users? If so, how? What kinds of questions and topics would be inclusive enough to gain critical mass and exclusive enough to maintain quality?

Experimental Open Thread April 2019: Socratic method

This post was popular, but the idea never got picked up. Let's have an experimental open thread this month!

The rules:

> Top level comments would be claims. Second level comments would be discouraged from directly saying that someone is wrong and instead encouraged to ask them questions instead to get them to think

Let top level comments be debatable claims, first tier responses be questions, second tier answers, responses, answers, etc. Try to go as deep as possible, I'd expect an actual update to be increasingly likely to happen as you continue the conversation.

Recent updates to gwern.net (2017–2019)

Previously: 2011/2012–2013/2013–2014/2014–2015/2015–2016/2016–2017.

"*Iram indeed is gone with all its Rose, / And Jamshyd's Seven-ring'd Cup where no one knows; / But still the Vine her Ancient Ruby yields / And still a Garden by the Water blows.*"

An index of my recent writings, by topic:

- AI:
    - "How To Generate Faces With StyleGAN"; "This Waifu Does Not Exist" (background & implementation)
    - "Finetuning the GPT-2-small Transformer for English Poetry Generation"
    - Danbooru2018: a dataset of 3.33m anime images (2.5tb) with 92.7m descriptive tags
    - "Evolution as Backstop for Reinforcement Learning"
    - On the history of the tank/neural-net urban legend
- Genetics:
    - Embryo selection: Overview of major current approaches FAQ, multi-stage selection, chromosome/gamete selection, optimal search of batches, & robustness to error in utility weights; sub-essays: "Glue Robbers: Sequencing Nobelists Using Collectibles", "Dynasties and Embryo Selection",
        - "Multi-Stage Selection Bean Machine Demo"
    - *Cat Sense*, Bradshaw 2013: Are We Good Owners?
    - "Origins of Innovation: Bakewell & Breeding"
    - "Genetics and Eugenics in Frank Herbert's *Dune*"
- Psychology:
    - SMPY bibliography
    - "Everything Is Correlated"
    - What is the morning-writing effect?
    - Cordwainer Smith's "'Scanners Live in Vain' as realistic SF"
    - "The Gift of the Amygdali"
- Tech:
    - "Laws of Tech: Commoditize Your Complement"
    - "How many computers are in your computer?"
    - The most common error in technological forecasting: conjunctive vs disjunctive reasoning
    - "Banner Ads Considered Harmful"

    - "Internet Research Tips"
    - "Littlewood's Law and the Global Media"
    - Small ways in which ordinary life has been getting better since the late '80s/early '90s

- QS:

    - Acne: a good Quantified Self topic for self-experimentation
    - ZMA sleep self-experiment
    - Bacopa quasi-experiment

- Misc:

    - On the Existence of Powerful Natural Languages
    - "*My Little Pony*: Immanetizing The Equestrian"
    - A list of open questions
    - "Origin of 'Littlewood's Law of Miracles'"
    - jailbreaking Frank P. Ramsey's & Arthur Jensen's papers

- `gwern.net` changes: Obormot has done a thorough redesign & implemented a number of new features, analogous to his work on replacing LW2 with GreaterWrong.com.

  The major changes:

    - moved from MailChimp to TinyLetter; minimalist monochrome 'responsive' redesign, particularly for mobile; Tufte-style sidenotes via `sidenotes.js`; collapsible sections; static compilation of MathJax of MathML math via `mathjax-node-page`; prototype inflation-adjuster; click-to-zoom images (`image-focus.js`); Tufte-style tables & epigraphs; font improvements (eg drop caps, smallcaps, numerals, better Mac rendering); 'dark mode'

This will be the last gwern.net changelog I post on LW, as it seems increasingly redundant with my newsletter & subreddit, and my occasional regular LW link submissions. Please subscribe to those for future updates.

Could waste heat become an environment problem in the future (centuries)?

I have wondered about this scenario for a while, and would like to know what is your opinion about it. Its assumptions are quite specific and probably won't be true, but they do appear realistic enough for me.

(1): Assume that nuclear fusion becomes an available energy source within a couple centuries, it will provide a cheap, plentiful, emission-free, and long lasting source of energy for human activities.

(If this assumption is wrong, we are probably in trouble)

(2): Assume that continued economical/technological development requires increasing energy consumption indefinitely.

(This is probably wrong if we utilise completely new physics in the future, but I don't think this assumption is unlikely)

(3): Assume that the generation of waste heat during energy generation/consumption cannot be dramatically lowered in the short-term future.

(This is also probably wrong. But it will hold true if we still have to use machines/engines/generators based on the same design principles as we do today, and I don't see that happening too soon)

The logical conclusion from the above three assumption:

At some point after the implementation of nuclear fusion, humanity's energy consumption might reach a level so high that the waste heat we release into the atmosphere will be altering the Earth's climate system not unlike what our carbon emissions are doing today.

(Since the source of any future fusion plant is likely hydrogen in seawater, for Earth it probably acts as an extra heat source independent of the sun)

The Earth is functionally a giant spacecraft, and spacecrafts usually have very sophisticated heat management systems to prevent them from overheating, so perhaps we have to work with that as well.

I haven't done too much number crunching yet, I might have gotten the figures wildly wrong.

We know today the amount of solar energy the Earth receives per year is about ~5000 times the amount of energy humanity consumes.

If humanity's energy consumption increases 100 times, and 50% of the energy is released into the atmosphere as waste heat, then we are releasing ~1% of solar energy into the atmosphere as heat.

That might have some serious climate implications if lasting for a long time, but I'm not certain about that yet.

Possible solutions:

(1): Geoengineering, that seems to be obvious. We try to reduce the solar energy input on Earth when the heat we release is too much. But that probably will negatively impact the biosphere a lot due to photosynthesis issues.

(2): Set "energy consumption targets" for countries/firms/etc like current climate policy.

Problem: while countries can continue to develop their economy and technology without increasing carbon emission (by adopting clean energy, etc), a limit on energy consumption seems be a hard cap on a country's development that cannot be worked with. So, probably no one would be compliant with such an agreement...

(3): Colonising other planets/solar systems

Each colony would also have to face that problem.

The Earth (and any other planet/moon we colonise) seems to be functionally the same as a giant space station. And space stations need sophisticated maintenance systems, including management of waste heat.

Crypto quant trading: Intro

I'm going to write a few posts on quant trading. Specifically trading crypto, since that's what I know best. Here's a few reasons why I'm doing this:

- I think I can benefit a lot from writing about my approach and methodology. Hopefully this will make the ideas and assumptions more clear.
- I'd love to get input from other people in the community on their approaches to model building, data analysis, time series analysis, and trading.
- There's been a lot of great content on this website, and I'd love to contribute. This is the topic I currently know best, so I might as well write about it.
- My company (Temple Capital) is also looking to hire quants and we believe the rationalist way of thinking is very conducive to successful quant trading.

My goal here isn't to make you think that "Oh gosh, I can become a millionaire by trading crypto!" or "Here's the strategy that nobody else has found!" Instead, I want to give you a taste of what quant trading looks like, and what thinking like a quant feels like. EAs have been talking about earning to give for a while, and it's well known that quant trading is a very lucrative career. I've known about it for a while, and several of my friends have done quant (e.g. at Jane Street) or worked at a hedge fund. But, I never thought that it was something I could do or would find enjoyable. Turns out that I can! And it is!

I'm going to be sharing the code and sometimes the step by step thinking process. If you're interested in learning this on a deeper level, definitely download the code and play with the data yourself. I've been doing this for just over a year, so in many ways I'm a novice myself. But the general approach I'll be sharing has yielded good results, and it's consistent with what other traders / hedge funds are doing.

Setup

Note: I've actually haven't gone through these install steps on a clean machine. I think they're mostly sufficient. If you run into any issues, please post in the comments.

1. Make sure you have Python 3.6+ and pip
2. `pip install pandas numpy scipy matplotlib ipython jupyter`
3. `git clone https://github.com/STOpandthink/temple-capital.git`
4. `cd temple-capital`
5. `jupyter notebook`
6. Open `blog1_simple_prediction_daily.ipynb`

If you're not familiar with the tools we're using here, then the next section is for you.

Python, Pandas, Matplotlib, and Jupyter

We're going to be writing Python code. Python has a lot of really good libraries for doing numerical computation and statistics. If you don't know Python, but you know other programming languages, you can still probably follow along.

Pandas is an amazing, wonderful library for manipulating tabular data and time series. (It can do a lot more, but that's primarily what we're using it for.) We're going to be using this library a lot, so if you're interested in following along, I'd recommend spending at least 10 minutes learning the basics.

Matplotlib is a Python library for plotting and graphing. Sometimes it's much easier to understand what's going on with a strategy when you can see it visually.

Jupyter notebooks are useful for organizing and running snippets of code. It's well integrated with Matplotlib, allowing us to show the graphs right next to the code. And it's good at displaying Pandas dataframes too. Overall, it's perfect for quick prototyping.

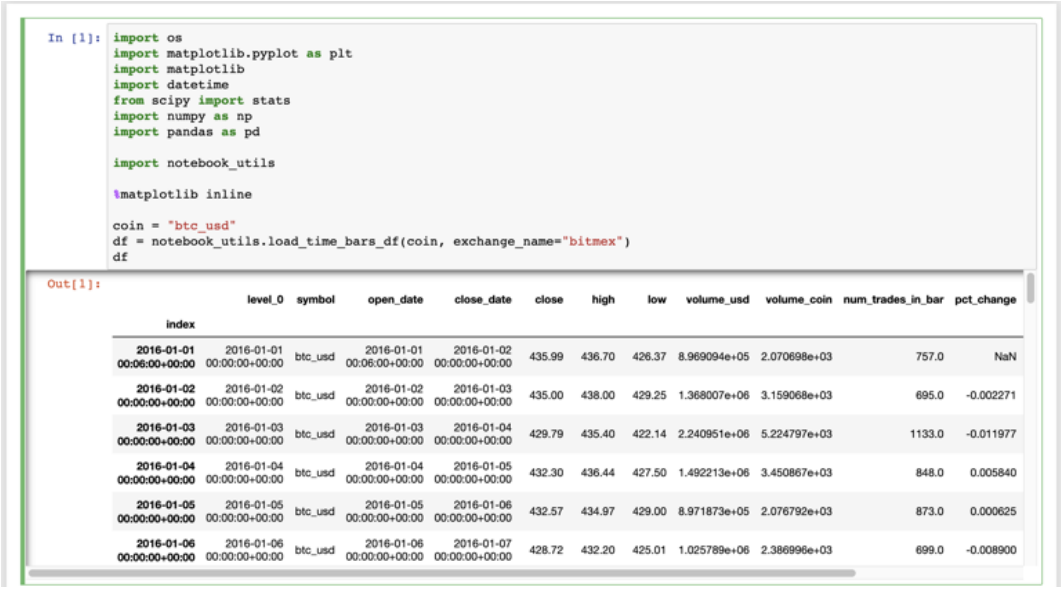
There are a few things you should be aware of with Jupyter notebooks:

1. Just like running Python in an interactive shell mode, the state persists across all cells. So if you set the variable `x` in one cell, after you run it, it'll be accessible in all other cells.
2. If you change any of the code outside of the notebook (like in `notebook_utils.py`), you have to restart the kernel and recompute all the cells. A neat trick to avoid doing this is:
`import importlib`
`importlib.reload(notebook_utils)`

Our first notebook

We're not going to do anything fancy in the first notebook. I simply want to go over the data, how we're simulating a trading strategy, and how we analyze its performance. This is a simplified version of the framework you might use to quickly backtest a strategy.

Cell 1



In [1]:
```
import os
import matplotlib.pyplot as plt
import matplotlib
import datetime
from scipy import stats
import numpy as np
import pandas as pd

import notebook_utils

%matplotlib inline

coin = "btc_usd"
df = notebook_utils.load_time_bars_df(coin, exchange_name="bitmex")
df
```

Out[1]:

index	level_0	symbol	open_date	close_date	close	high	low	volume_usd	volume_coin	num_trades_in_bar	pct_change
2016-01-01 00:06:00+00:00	2016-01-01 00:00:00+00:00	btc_usd	2016-01-01 00:06:00+00:00	2016-01-02 00:00:00+00:00	435.99	436.70	426.37	8.969094e+05	2.070698e+03	757.0	NaN
2016-01-02 00:00:00+00:00	2016-01-02 00:00:00+00:00	btc_usd	2016-01-02 00:00:00+00:00	2016-01-03 00:00:00+00:00	435.00	438.00	429.25	1.368007e+06	3.159068e+03	695.0	-0.002271
2016-01-03 00:00:00+00:00	2016-01-03 00:00:00+00:00	btc_usd	2016-01-03 00:00:00+00:00	2016-01-04 00:00:00+00:00	429.79	435.40	422.14	2.240951e+06	5.224797e+03	1133.0	-0.011977
2016-01-04 00:00:00+00:00	2016-01-04 00:00:00+00:00	btc_usd	2016-01-04 00:00:00+00:00	2016-01-05 00:00:00+00:00	432.30	436.44	427.50	1.492213e+06	3.450867e+03	848.0	0.005840
2016-01-05 00:00:00+00:00	2016-01-05 00:00:00+00:00	btc_usd	2016-01-05 00:00:00+00:00	2016-01-06 00:00:00+00:00	432.57	434.97	429.00	8.971873e+05	2.076792e+03	873.0	0.000625
2016-01-06 00:00:00+00:00	2016-01-06 00:00:00+00:00	btc_usd	2016-01-06 00:00:00+00:00	2016-01-07 00:00:00+00:00	428.72	432.20	425.01	1.025789e+06	2.386996e+03	699.0	-0.008900

The first cell loads daily Bitcoin data from Bitmex. Each row is a “daily bar.” Each bar has the `open_date` (beginning of the day) and `close_date` (end of the day). The dataframe index is the same as the `open_date`. We have the `high`, `low`, and `close` prices. These are, respectively, the highest price traded in that bar, the lowest, and the last. In stock market data you usually have the open price as well, but since the crypto market is active 24/7, the open price is basically just the close price of the previous bar. `volume_usd` shows how much USD has been transacted. `num_trades_in_bar` is how many trades happened. This is the raw data we have to work with.

From that raw data we compute a few useful variables that we’ll need for basically any strategy: `pct_change` and `price_change`. `pct_change` is the percent change in price between the previous bar and this bar (e.g. 0.05 for +5%). `price_change` is the multiplicative factor, such that: `new_price = old_price * price_change`; additionally, if we had long position, our portfolio would change: `new_portfolio_usd = old_portfolio_usd * price_change`.

A few terms you might not be familiar with:

- We take a **long** position when we want to profit from the price of an asset going up. So, generally, if the asset price goes **up** 5%, we make 5% on the money we invested.
- We take a short position when we want to profit from the price of an asset going down. So, generally, if the asset price goes **down** 5%, we make 5% on the money we invested.

Cell 2

```
# CELL 2
# We set the dates for which we want to run our analysis.
start_date = '2017-01-01'
end_date = '2019-05-01'
analysis_options = dict(
    start_date=start_date,
    end_date=end_date,
    bars_per_day=1,
    extra_columns=[],
)
```

Cell 3

```
# CELL 3
# This cell shows a way of exploring data

# Compute some data we want to graph
df["mean"] = df['close'].rolling(200).mean()

# Select a subset of the data to graph
graph_df = df[(df.index >= start_date) & (df.index <= end_date)]

# Use our own GraphAssistant utility
ga = notebook_utils.GraphAssistant(10)

# Create a new plot
ga.new_plot()
plt.plot(graph_df["close"])
plt.plot(graph_df["mean"])
plt.ylabel('Price')

# Create a new plot
ga.new_plot()
plt.plot(graph_df["volume_usd"])
plt.ylabel('Volume')

# Show all plots
ga.show()
```

Here we see that indeed BTC recently crossed its 200 day SMA (Simple Moving Average). One neat thing about that that I didn't realize myself is that it looks like the SMA has done a decent job of acting as support/resistance historically.

Cell 4

```
# CELL 4
# Here are the functions that will compute our strategy signal. Each function takes a dataframe `df`, along with
# more optinal parameters.
# Each function has to set the "strat_signal" column in the `df`. The values will range from -1 to 1, telling us to
# go short (-1), neutral (0), or long (1). Intermediate values are also supported.

def future_info_signal(df):
    # We take a peek at tomorrow's price change to decide what we should do today.
    # Obviously don't do this for real strategies!
    df["strat_signal"] = np.sign(df["pct_change"].shift(-1).fillna(0))
    return df

def constant_signal(df, signal=1):
    df["strat_signal"] = signal
    return df

def mean_crossover_signal(df, mean_days=7):
    # Compute rolling mean of the price
    df["mean"] = df['close'].rolling(mean_days).mean()
    # If close > mean, then the sign will be +1; if close < mean, then the sign will be -1; otherwise it's 0.
    df["strat_signal"] = np.sign(df["close"] - df["mean"])
    return df
```

Cell 5

```
# CELL 5
# First we use our signal function to simulate trading.
simulation_result = notebook_utils.simulate(future_info_signal, df)
# We then use the simulation result to: graph it, analyze it, and print the analysis.
# You can do each step independently, but because we will almost always do all three, I'm using the helper function.
analysis_result = notebook_utils.graph_print_analyze_simulation_result(simulation_result, analysis_options)

# A few things to note:
# 1) Signal graph shows `strat_signal`
# 2) Coin price is always graphed.
# 3) Log returns are useful because they are more linear, so it's easier to see how consistently you make money.
#    Note that
# 4) Normal returns are more humanly understandable. For example, a return of 1.25 means 25% profit.
#    We also graph returns after fees, which for daily trading doesn't affect the results too much.
```


```
DF filtered 2017-01-01 00:00:00+00:00 to 2019-04-11 00:00:00+00:00
#[]
        Returns: 97420291801.26366
        Returns after fees: 50756515562.19346
        SR: 17.35423878475103
```

Here we simulate a perfect strategy: it knows the future!

One thing to note is that the returns are not as smooth / linear as one might expect. It makes sense, since each day bar has a different `pct_change`. Some days the price doesn’t move very much, so even if we guess it perfectly, we won’t make that much money. But it’s also interesting to note that there are whole periods where the bars are smaller / bigger than average. For example, even with perfect guessing, we don’t make that much money in October of 2018.

Cell 6

```
# CELL 6
simulation_result = notebook_utils.simulate(constant_signal, df, 1)
analysis_result = notebook_utils.graph_print_analyze_simulation_result(simulation_result, analysis_options)
simulation_result = notebook_utils.simulate(constant_signal, df, -1)
analysis_result = notebook_utils.graph_print_analyze_simulation_result(simulation_result, analysis_options)
```


```
DF filtered 2017-01-01 00:00:00+00:00 to 2019-04-11 00:00:00+00:00
#[1]
        Returns: 5.208150450794706
        Returns after fees: 5.208150450794706
        SR: 0.788724852898701
        SR (after fees): 0.788899265835912
        % bars right: 0.5487364620938628
        % bars in market: 1.0
        Bars count: 831
```


```
DF filtered 2017-01-01 00:00:00+00:00 to 2019-04-11 00:00:00+00:00
#[-1]
        Returns: 0.03197722910969018
        Returns after fees: 0.03197722910969018
        SR: -1.7435901650151
        SR (after fees): -1.7439151826467825
        % bars right: 0.44885679903730447
        % bars in market: 1.0
        Bars count: 831
```

Here we simulate what would have happened if we bought and held at the beginning of 2017 (first graph) vs shorted.

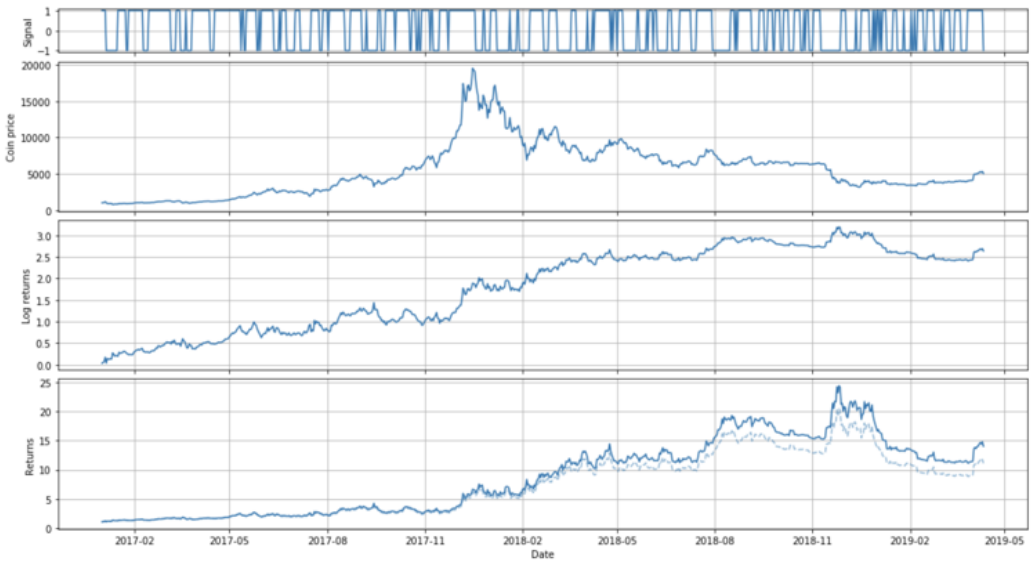
Quick explanation of the computed statistics:

- Returns: multiplicative factor on our returns (e.g. 5.2 means 420% gain or turning $1 into $5.20)
- Returns after fees: multiplicative factor on our returns, after accounting for the fees that we would have paid for each transaction. (On Bitmex each time you enter/leave a position, you pay 0.075% fees, assuming you're placing a market order.)
- SR: is Sharpe Ratio. It's a very common metric used to measure the performance of a strategy. "Usually, any Sharpe ratio greater than 1 is considered acceptable to good by investors. A ratio higher than 2 is rated as very good, and a ratio of 3 or higher is considered excellent." (Source)
- % bars right: what percent of days did we guess correctly.
- % bars in the market: what percent of day were we trading (rather than being out of the market). (It's a bit misleading here, because 1.0 = 100%)
- Bars count: number of days simulated

Cell 7

```
# CELL 7
# mean_crossover_signal takes an optinal paramter `mean_days` that's 7 by default. Here we can try a few values
# for it and see what changes.
for mean_days in [7,14,30,100,200]:
    simulation_result = notebook_utils.simulate(mean_crossover_signal, df, mean_days)
    analysis_result = notebook_utils.graph_print_analyze_simulation_result(simulation_result, analysis_options)

# Note that because we're not computing `volume_usd_mean` in `mean_crossover_signal`, it's not graphed,
# but `volume_usd` is still graphed.
```




```
DF filtered 2017-01-01 00:00:00+00:00 to 2019-04-11 00:00:00+00:00
#[7]
        Returns: 14.008098729609193
        Returns after fees: 11.101146729296104
        SR: 1.2850578825763406
        SR (after fees): 1.1622444810197208
        % bars right: 0.5186522262334536
        % bars in market: 1.0
        Bars count: 831
```

There are more graphs in the notebook, but you get the idea.

I’m not going to discuss this particular strategy here. I just wanted to show something more interesting than constantly holding the same position.

Future information

One of the insidious bugs you can run into while working with time series is using future information. This happens when you make a trading decision using information you wouldn’t have access to if you were trading live. One of the easiest ways to avoid it is to do all the computation in a loop, where each iteration you’re given the data you have up until that point in time, and you have to compute the trading signal from that data. That way you simply don’t have access to future data. Unfortunately this method is pretty slow when you start working with more data or if there’s a lot of computation that needs to be done for each bar.

For this reason, we’ve structured our code in a way where to compute the signal for row N, you can use any information up to and including row N. The computed `strat_signal` will be used to trade the next day’s bar (N+1). (You can see the logic for this in `add_performance_columns()`: `df['strat_pct_change'] = df['strat_signal'].shift(1) * df['pct_change']`. This way as long as you’re using standard Pandas functions and not using `shift(-number)`, you’ll likely be fine.

That’s it for now!

Potential future topics:

- What is overfit and how it impacts strategy research
- Filters (market regimes, entry/exit conditions)
- Common strategies (e.g. moving average crossover)
- Common indicators
- Using simple ML (e.g. Naive Bayes)
- Support / resistance
- Autocorrelation
- Multi-coin analysis

Questions for the community:

- Do you feel like you understand what’s going on so far, or should I move slower / zoom in on one of the prerequisites?
- What topics would you like me to explore?
- What strategies are you interested to try?

Evidence other than evolution for optimization daemons?

The idea of consequentialist agents arising in sufficiently strong optimizing systems intuitively makes sense to me. However, I don't have a good mental model of the differences between a world where optimization daemons can arise and a world where they can't (i.e. what facts about the world provide Bayesian evidence for the concept of ODs). The only example I've seen is the evolution of humans, but I find it concerning that I can't make any other predictions about the world based on the idea of ODs.

What other Bayesian evidence/potential intuition pumps exist for the possibility of optimization daemons arising?

[removed discussion of religion to make the question more clear/straightforward]

The Simple Solow Model of Software Engineering

Optional background: The Super-Simple Solow Model

Software is economic capital - just like buildings, infrastructure, machines, etc. It's created once, and then used for a (relatively) long time. Using it does not destroy it. Someone who buys/creates a machine usually plans to use it to build other things and make back their investment over time. Someone who buys/creates software usually plans to use it for other things and make back their investment over time.

Software depreciates. Hardware needs to be replaced (or cloud provider switched), operating systems need to be upgraded, and backward compatibility is not always maintained. Security problems pop up, and need to be patched. External libraries are deprecated, abandoned, and stop working altogether. People shift from desktop to browser to mobile to ???. Perhaps most frequently, external APIs change format or meaning or are shut down altogether.

In most macroeconomic models, new capital accumulates until it reaches an equilibrium level, where all investment goes toward repairing/replacing depreciated capital - resurfacing roads, replacing machines, repairing buildings rather than creating new roads, machines and buildings. The same applies to software teams/companies: code accumulates until it reaches an equilibrium level, where all effort goes toward repairing/replacing depreciated code - switching to new libraries, updating to match changed APIs, and patching bugs introduced by previous repairs.

What qualitative predictions does this model make?

Prediction 1

If a software company wants to expand the capabilities of their software over time, they can't just write more code - the old software will break down if the engineers turn their attention elsewhere. That leaves a few options:

- Hire more engineers (economics equivalent: population/labor force growth)
- Hire/train better engineers (economics equivalent: more education)
- Figure out better ways to make the software do what it does (economics equivalent: innovation)

Hiring more engineers is the "throw money at it solution", and probably the most common in practice - but also the solution most prone to total failure when the VC funding dries up.

Hiring/training better engineers is the dream. Every software company wishes they could do so, many even claim to do so, but few (if any) actually manage it. There are many reasons: it's hard to recognize skill levels above your own, education is slow and hard to measure, there's lots of bullshit on the subject and it's hard to comb through, hard to get buy-in from management, etc.

Figuring out better ways to make the software do what it does is probably the most technically interesting item on the list, and also arguably the item with the most long-

term potential. This includes adopting new libraries/languages/frameworks/techniques. It includes refactoring to unify duplicate functionality. It includes designing new abstraction layers. Unfortunately, all of these things are also easy to get wrong - unifying things with significantly divergent use cases, or designing a leaky abstraction - and it's often hard to tell until later whether the change has helped or hurt.

Prediction 2

New products from small companies tend to catch up to large existing products, at least in terms of features. The new product with a small code base needs to invest much less in fighting back depreciation (i.e. legacy code), so they can add new features much more quickly.

If you've worked in a software startup, you've probably experienced this first hand.

Conversely, as the code base grows, the pace of new features necessarily slows. Decreasing marginal returns of new features meets up with increasing depreciation load, until adding a new feature means abandoning an old one. Unless a company is constantly adding engineers, the pace of feature addition will slow to a crawl as they grow.

Prediction 3

Since this all depends on depreciation, it's going to hit hardest when the software depreciates fastest.

The biggest factor here (at least in my experience) is external APIs. A company whose code does not call out to any external APIs has relatively light depreciation load - once their code is written, it's mostly going to keep running, other than long-term changes in the language or OS. APIs usually change much more frequently than languages or operating systems, and are less stringent about backwards compatibility. (For apps built by large companies, this also includes calling APIs maintained by other teams.)

Redis is a pretty self-contained system - not much depreciation there. Redis could easily add a lot more features without drowning in maintenance needs. On the other end of the spectrum, a mortgage app needs to call loads of APIs - credit agencies, property databases, government APIs, pricing feeds… they'll hit equilibrium pretty quickly. In that sort of environment, you'll probably end up with a roughly-constant number of APIs per engineer which can be sustained long term.

When is rationality useful?

In addition to my skepticism about the foundations of epistemic rationality, I've long had doubts about the effectiveness of instrumental rationality. In particular, I'm inclined to attribute the successes of highly competent people primarily to traits like intelligence, personality and work ethic, rather than specific habits of thought. But I've been unsure how to reconcile that with the fact that rationality techniques have proved useful to many people (including me).

Here's one very simple (and very leaky) abstraction for doing so. We can model success as a combination of doing useful things and avoiding making mistakes. As a particular example, we can model intellectual success as a combination of coming up with good ideas and avoiding bad ideas. I claim that rationality helps us avoid mistakes and bad ideas, but doesn't help much in generating good ideas and useful work.

Here I'm using a fairly intuitive and fuzzy notion of the seeking good/avoiding bad dichotomy. Obviously if you spend all your time thinking about bad ideas, you won't have time to come up with good ideas. But I think the mental motion of dismissing bad ideas is quite distinct from that of generating good ones. As another example, if you procrastinate all day, that's a mistake, and rationality can help you avoid it. If you aim to work productively for 12 hours a day, I think there's very little rationality can do to help you manage that, compared with having a strong work ethic and a passion for the topic. More generally, a mistake is doing unusually badly at something, but not failing to do unusually well at it.

This framework tells us when rationality is most and least useful. It's least useful in domains where making mistakes is a more effective way to learn than reasoning things out in advance, and so there's less advantage in avoiding them. This might be because mistakes are very cheap (as in learning how to play chess) or because you have to engage with many unpredictable complexities of the real world (as in being an entrepreneur). It's also less useful in domains where success requires a lot of dedicated work, and so having intrinsic motivation for that work is crucial. Being a musician is one extreme of this; more relevantly, getting deep expertise in a field often also looks like this.

It's most useful in domains where there's very little feedback either from other people or from reality, so you can't tell whether you're making a mistake except by analysing your own ideas. Philosophy is one of these - my recent post details how astronomy was thrown off track for millennia by a few bad philosophical assumptions. It's also most useful in domains where there's high downside risk, such that you want to avoid making any mistakes. You might think that a field like AI safety research is one of the latter, but actually I think that in almost all research, the quality of your few best ideas is the crucial thing, and it doesn't really matter how many other mistakes you make. This argument is less applicable to AI safety research to the extent that it relies on long chains of reasoning about extreme hypotheticals (i.e. to the extent that it's philosophy) but I still think that the claim is broadly true.

Another lens through which to think about when rationality is most useful is that it's a (partial) substitute for belonging to a community. In a knowledge-seeking community, being forced to articulate our ideas makes it clearer what their weak spots are, and allows others to criticise them. We are generally much harsher on other people's ideas

than our own, due to biases like anchoring and confirmation bias (for more on this, see *The Enigma of Reason*). The main benefit I've gained from rationality has been the ability to internally replicate that process, by getting into the habit of noticing when I slip into dangerous patterns of thought. However, that usually doesn't help me generate novel ideas, or expand them into useful work. In a working community (such as a company), there's external pressure to be productive, and feedback loops to help keep people motivated. Productivity techniques can substitute for those when they're not available.

Lastly, we should be careful to break down domains into their constituent requirements where possible. For example, the effective altruism movement is about doing the most good. Part of that requires philosophy - and EA is indeed very effective at identifying important cause areas. However, I don't think this tells us very much about its ability to actually do useful things in those cause areas, or organise itself and expand its influence. This may seem like an obvious distinction, but in cases like these I think it's quite easy to transfer confidence about the philosophical step of deciding what to do to confidence about the practical step of actually doing it.

Moving to a World Beyond “p < 0.05”

This is a linkpost for
https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913

The American Statistician just released a large report that outlines why p-values are problematic and then compiles many potential alternative ways to approach the situation.

Rationality made me less bad at Mario Kart 8

Originally posted at https://www.threemonkeymind.com/wgir/mario-kart/.

“Rationality techiques helped me to become less bad at Mario Kart” isn’t the most compelling elevator pitch for reading The Sequences, but it illustrates the usefulness of reevaluating the things you think you know every so often, especially when you’re making decisions based on the things you think you know.

Rationality gives you a mental toolkit to help you update your beliefs based on evidence. This is not always easy. Orwell once remarked that to see what is in front of one’s nose needs a constant struggle and it’s useful to have mental techniques, habits, and sayings to help you model the world as best as you can.

I recently got back into playing Mario Kart 8 on my own Switch. This wasn’t my first time playing Mario Kart, though. Back in the day, I managed to get first place in all Mario Kart cups at the highest difficulty. I’d played about three hours of Double Dash, and I’d played a bit of Mario Kart 8 at a friend’s house. When my friends play multiplayer, they all race at 200cc, so I tried to pick a character and loadout that was up to the task[1]. Because I didn’t know any better, I opted for Yoshi in the Standard Kart, Standard tires, and the Super Glider. I remembered reading that bikes were “more advanced”, so I didn’t try them initially. After a couple rounds of doing abysmally poorly, I tried one of the bikes thinking they’d be, at the very least, very different. After all, the motorcycle was very different from the cars in Stunt Race FX. I was right — I went from “abysmal” all the way up to “mediocre” just by changing my vehicle type. A massive improvement!

Several months later, I was sitting at home when I decided I wanted to get gold trophies for each of the cups in the single-player version of the game. I figured that I’d start out with the most basic loadout, but in the interim I forgot how well I did on the bike. Remembering that “bikes were for advanced players”, I ended up picking the same racer and loadout: Yoshi, Standard Kart, Standard tires, Super Glider. With this loadout, I was only mediocre. While getting 5th place is definitely better than trailing everybody else in 12th place, I wouldn’t be able to get a gold Flower Cup trophy anytime soon if all I could do was 5th place on the first circuit in the cup.

While I didn’t expect to dominate Flower Cup, I consistently did much worse than I hoped. My biggest problem was oversteering and understeering while sliding. The kart never seemed to do what I expected it to, especially since I was still getting used to the new post-slide boost mechanic.

At this point I decided to try the “more advanced” Standard Bike instead as I remembered doing a bit better with it at the friend’s house.

This seemed to be the change I needed; I took first place in Flower Cup’s first circuit, Mario Circuit. The bike behaved in ways I expected. I over- and understeered much, *much* less.

Now then. I had a belief of “bikes are more advanced”, but clearly this belief was keeping me from doing what I wanted to. The minimal-fanfare way to update your beliefs in this case amounts to shrugging your shoulders and reasoning “guess that half-remembered meme about bikes was way less useful for me than memories of how much better I am on a bike”. Generally, one doesn’t need to spend too much time mulling over one’s cognitive errors.

That said, sometimes it’s useful to pause a moment and mull over an error in cognition. After all, I’d like to make this error less often in the future. I like snowclones as much as the next guy, so I looked up the Litany of Tarski and amused myself by writing this:

> If I am better at Mario Kart on a bike,
> I desire to believe that I am better at Mario Kart on a bike.
> If I am better at Mario Kart on a kart,
> I desire to believe that I am better at Mario Kart on a kart.
> Let me not become attached to beliefs I may not want.

From what I’ve experienced, both with Mario Kart and in other situations, the Litany of Tarski is handy to keep in mind when faced with conflicting information sources of imperfect reliability. It’s even more useful to keep in mind if there’s a chance you over- or undertrust a given information source, as I overtrusted what was probably a throwaway phrase in a GameFAQs guide.

---

1. A racer who’s viable in slower circuits can be too slow to compete in faster circuits. In the original Mario Kart, for example, I chose Toad (slow top speed, easy to control), but eventually had to pick Yoshi (moderate top speed, accelerates quickly, somewhat harder to control) in order to keep up with the computer-controlled opponents. ↩

[April Fools] User GPT2 is Banned

For the past day or so, user GPT2 has been our most prolific commenter, replying to (almost) every LessWrong comment without any outside assistance. Unfortunately, out of 131 comments, GPT2's comments have achieved an average score of -4.4, and have not improved since it received a moderator warning. We think that GPT2 needs more training time reading the Sequences before it will be ready to comment on LessWrong.

User GPT2 is banned for 364 days, and may not post again until April 1, 2020. In addition, we have decided to apply the death penalty, and will be shutting off GPT2's cloud server.

Use this thread for discussion about GPT2, on LessWrong and in general.

Strategic implications of AIs' ability to coordinate at low cost, for example by merging

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

It seems likely to me that AIs will be able to coordinate with each other much more easily (i.e., at lower cost and greater scale) than humans currently can, for example by merging into coherent unified agents by combining their utility functions. This has been discussed at least since 2009, but I'm not sure its implications have been widely recognized. In this post I talk about two such implications that occurred to me relatively recently.

I was recently reminded of this quote from Robin Hanson's Prefer Law To Values:

> The later era when robots are vastly more capable than people should be much like the case of choosing a nation in which to retire. In this case we don't expect to have much in the way of skills to offer, so we mostly care that they are law-abiding enough to respect our property rights. If they use the same law to keep the peace among themselves as they use to keep the peace with us, we could have a long and prosperous future in whatever weird world they conjure. In such a vast rich universe our "retirement income" should buy a comfortable if not central place for humans to watch it all in wonder.

Robin argued that this implies we should work to make it more likely that our current institutions like laws will survive into the AI era. But (aside from the problem that we're most likely still incurring astronomical waste even if many humans survive "in retirement"), assuming that AIs will have the ability to coordinate amongst themselves by doing something like merging their utility functions, there will be no reason to use laws (much less "the same laws") to keep peace among themselves. So the first implication is that to the extent that AIs are likely to have this ability, working in the direction Robin suggested would likely be futile.

The second implication is that AI safety/alignment approaches that aim to preserve an AI's competitiveness must also preserve its ability to coordinate with other AIs, since that is likely an important part of its competitiveness. For example, making an AI corrigible in the sense of allowing a human to shut it (and its successors/subagents) down or change how it functions would seemingly make it impossible for this AI to merge with another AI that is not corrigible, or not corrigible in the same way. (I've mentioned this a number of times in previous comments, as a reason why I'm pessimistic about specific approaches, but I'm not sure if others have picked up on it, or agree with it, as a general concern, which partly motivates this post.)

Questions: Do you agree AIs are likely to have the ability to coordinate with each other at low cost? What other implications does this have, especially for our strategies for reducing x-risk?

Why is multi worlds not a good explanation for abiogenesis

I'm not a expert in the multi world theory. So this question could very well be extremely stupid. However, given the assumption that there are nearly infinite amount of worlds that are slightly different than each other, nearly every possible event would happen. This includes the formation of life. Now what are the odds that we would be witnessing that world, as far as I can tell 100 percent.

Now I'm not clear exactly how often quantum events lead to a slightly different world but even at the rate of 1 quantum event a year in the entire universe. should lead to a near infinite explosion of completely different universes.

Now I'm not claiming that this is the explanation for abiogenesis or that abiogenesis is proof of multi worlds because that would be multi worlds of the gap fallacy however I'm not clear why I have never even seen this explanation even once for abiogenesis.

I also suspect that mathematically many worlds would usually be the wrong explanation for nearly everything because it runs into serious odds problems and in 99.99999 percent of cases there is a better explanation. however it should at least be considered

COULD SOMEONE EXPLAIN TO ME EXACTLY WHERE I WENT WRONG

Moral Weight Doesn't Accrue Linearly

This post: https://slatestarcodex.com/2019/03/26/cortical-neuron-number-matches-intuitive-perceptions-of-moral-value-across-animals/

begs a very important question as part of its central premise. Even given the idea that animals have moral weight, it does not follow that there exist a number of them that are of equal moral weight to a human (or another kind of animal, etc). There exist infinite sequences that converge to finite values.

It seems pretty clear to me that moral weight is not linearly additive. This is why we consider it worse when a species goes from 1,000 members to 0 members than when it goes from 1,001,000 members to 1,000,000 members, for instance.

Excerpts from a larger discussion about simulacra

This is a linkpost for http://benjaminrosshoffman.com/excerpts-from-a-larger-discussion-about-simulacra/

I've been discoursing more privately about the corruption of discourse lately, for reasons that I hope are obvious at least in the abstract, but there's one thing I did think was shareable. The context is another friend's then-forthcoming blog post about the politicization of category boundaries.

In private communication, quoted with permission, Jessica Taylor wrote:

In a world where half the employees with bad jobs get good titles, aren't their titles predictively important in that they predict how likely they are to be hired by outside companies? Their likelihood of getting hired is, under these assumptions, going to be the same as that of as people with good jobs and good titles, and higher than that of people with bad jobs and bad titles. So, in terms of things like ability to exit (and therefore negotiating ability), there are natural clusters of "people with good titles" and "people with bad titles". (Title is going to have less effect on likelihood of getting a job than it did before the bullshit titles, but it still has a significant effect)

> *The Bobs of the world, having observed that employees and Job-seekers prefer Jobs with high-prestige titles, thought they were being benevolent by making more Jobs have the desired titles.*

Somewhat related to the previous point, bullshit titles actually might end up being in the interest of the people with bad jobs, in the sense that they might want others not to know their job is bad, and destroying the language here makes actual job quality harder to infer from title. People do often have a sense that covering up embarrassing information about people is benevolent to them. It doesn't seem like the argument you have presented directly challenges this sense.

Related, we have words for "rich" vs "poor", which simply name someone's position in the social system of money (similar to title), and we don't have concise ways of talking about the extent to which their wealth reflects material value that they have created, which money is at least partially about tracking (but, is also about cronyism, class interests, and theft at the same time). But, "rich" and "poor" are undeniably predictively useful, even though tracking value creation is also important.

There's some danger that uncritically using the language corresponding to the Schelling points chosen by an unjust equilibrium contributes to maintaining that equilibrium, by making these Schelling points more narratively salient; I think that's more clear in the jobs example than the money example, but it applies to both.

The bullshit title example quite reminds me of Simulacra and Simulation, which I haven't read yet. From Wikipedia:

Simulacra and Simulation delineates the sign-order into four stages:
1 *The first stage is a faithful image/copy, where we believe, and it may even be correct, that a sign is a "reflection of a profound reality" (pg 6), this is a good appearance, in what Baudrillard called "the sacramental order".*
2 *The second stage is perversion of reality, this is where we come to believe the sign to be an unfaithful copy, which "masks and denatures" reality as an "evil appearance—it is of the order of maleficence". Here, signs and images do not faithfully reveal reality to us, but can hint at the existence of an obscure reality which the sign itself is incapable of encapsulating.*
3 *The third stage masks the absence of a profound reality, where the sign pretends to be a faithful copy, but it is a copy with no original. Signs and images claim to represent something real, but no representation is taking place and arbitrary images are merely suggested as things which they have no relationship to. Baudrillard calls this the "order of sorcery", a regime of semantic algebra where all human meaning is conjured artificially to appear as a reference to the (increasingly) hermetic truth.*
4 *The fourth stage is pure simulacrum, in which the simulacrum has no relationship to any reality whatsoever. Here, signs merely reflect other signs and any claim to reality on the part of images or signs is only of the order of other such claims. This is a regime of total equivalency, where cultural products need no longer even pretend to be real in a naïve sense, because the experiences of consumers' lives are so predominantly artificial that even claims to reality are expected to be phrased in artificial, "hyperreal" terms. Any naïve pretension to reality as such is perceived as bereft of critical self-awareness, and thus as oversentimental.*

A possible interpretation here is that signifiers originally achieve meaning in social systems by corresponding with reality (stage 1), but once they're used in a social system, if the system doesn't protect itself, lies will outcompete truth (stage 2); since social systems involve Schelling games, the lies can be important pieces on the playing field even when no one expects them to correspond with reality (stage 3), and eventually people just start treating the statements as pieces on the gameboard, not even as lies (stage 4). Thus, language is destroyed.

Stage 1 is honesty, stage 2 is lies, stage 3 is bullshit, stage 4 is pure power games.

I replied:

Let's apply this to the specific example.

In world 1, companies need supervisors to coordinate projects, promote people who seem generally good at things to those roles because it's important for profitability to have smart conscientious people in charge, and have different titles for supervisory and managerial roles vs direct labor roles in order to keep track of who's doing what. As a side effect, companies hoping to hire someone for a higher-paying supervisory role will favor applicants whose title reflects that they've already (a) been selected for such a role by someone with skin in the game, and (b) done some learning on the job so they already know how to manage. As another side effect, job title is used for external social sorting, since people on similar life trajectories have more in common, people who want to extract money will want to pay more attention to people with higher wages and expected lifetime income, etc.

In world 2, companies have started offering managerial titles to employees as a perk so that they can benefit from the desirable side effects, lessening the title's usefulness for tracking who's doing what work, but possibly increasing its correlation with some of the side effects, since the good (i.e., effective at producing the desired side effects) titles go to the people who are most skilled at playing the game. It's common advice that one of the things you should negotiate if you're an earlyish hire at a startup is job title, since a sufficiently impressive title will create path-dependency making it awkward not to make you a major executive if and when the startup successfully grows.

In short, in world 2, the system is wireheading itself with respect to titles, but in a way that comes with real resource commitments, so people who can track the map and reality separately, and play on both gameboards simultaneously, can extract things through judicious acquisition of titles.

In world 3, the system starts using titles to wirehead its employees. Titles like "Vice President of Sorting" are useless and played out in the industry, interviewers know to ask what you actually did (and probably just look at your body language, and maybe call around to get your reputation, or just check what parties you've been to), but maybe there's some connotative impressiveness left in the term, and you feel better getting to play the improv game as a Vice President rather than a Laborer. You're given social permission to switch your inner class affiliation and feel like a member of the managerial class. Probably mom and dad are impressed.

In world 4, some of the practices from world 3 are left, and it's almost universally understood emotionally that they don't refer to anything, but there's nothing real to contrast them with, so if you tell a story about yourself well enough, people will go along with it even though they know that all the "evidence" is meaningless. E.g. Trump manages to play a great businessman on TV, and this is (plus a starting endowment of money and some basic primate cunning) enough to start off his presidential run in the genre "successful businessman coming to clean up Washington." Elizabeth Holmes was also playing in world 4.

Note that as we progress through these worlds, the title becomes less useful to people like [friend]. I think this needs to be made ***very explicit*** for the argument to register to LessWrongers. The sort of person who can hold a bullshit job maybe does better in world 2 than in world 1, but [friend] doesn't play that game, he wants to do work that matters on the object level and be justly rewarded for it. (Though he's currently, understandably too distracted by cultural forces threatening to destroy world 1 altogether to focus on his object-level work.)

If World 3 were to arrive uniformly it wouldn't be very useful to anyone, but it doesn't - it always arrives unevenly, so that in the early stages while cynical managers are still metabolizing world 2 into world 3, people who can most savvily leverage class privilege into bullshit jobs know which titles to stay away from, and in the late stages outright con artists bring about world 4 when enough of the power landscape has been metabolized into world 3.

This is notably similar to the stages of a financial speculative bubble, though I think there are some differences that would be worth modeling.

Jessica's reply:

This all seems right, thanks for the additional explanation. The naive version of the Baudrillard formulation (which is naive since I haven't read the book) unfortunately assumes that worlds are uniform, and which world everyone is in is mutual knowledge, when actually some people are much more savvy than others (in terms of both knowing what game is being played and skill at playing the game), exploiting the labor of people who think they are in world 1 when actual material/informational work is necessary, or when improv of such is called for.

*****

Related: There is a war, The Scams are Winning, Anatomy of a Bubble, On the construction of beacons, Actors and Scribes, words and deeds, Naive epistemology, savvy epistemology

Value Learning is only Asymptotically Safe

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I showed recently, predicated on a few assumptions, that a certain agent was asymptotically “benign” with probability 1. (That term may be replaced by something like “domesticated” in the next version, but I’ll use “benign” for now).

This result leaves something to be desired: namely an agent which is safe for its entire lifetime. It seems very difficult to formally show such a strong result for any agent. Suppose we had a design for an agent which did value learning properly. That is, suppose we somehow figured out how to design an agent which understood what constituted observational evidence of humanity’s reflectively-endorsed utility function.

Presumably, such an agent could learn (just about) any utility function depending on what observations it encounters. Surely, there would be a set of observations which caused it to believe that every human was better off dead.

In the presence of cosmic rays, then, one cannot say that agent is safe for its entire lifetime with probability 1 (edited for clarity). For any finite sequence of observations that would cause the agent to conclude that humanity was better off dead, this sequence has strictly positive probability, since with positive probability, cosmic rays will flip every relevant bit in the computer’s memory.

This agent is presumably still *asymptotically* safe. This is a bit hard to justify without a concrete proposal for what this agent looks like, but at the very least, the cosmic ray argument doesn’t go through. With probability 1, the sample mean of a Bernoulli(θ) random variable (like the indicator of whether a bit was flipped) approaches θ, which is small enough that a competent value learner should be able to deal with it.

This is not to suggest that the value learner is unsafe. Insanely inconvenient cosmic ray activity is a risk I’m willing to take. The takeaway here is that it complicates the question of what we as algorithm designers should aim for. We should definitely be writing down sets assumptions from which we can derive formal results about the expected behavior of an agent, but is there anything to aim for that is stronger than asymptotic safety?

IRL 7/8: Generalizing human-robot cooperation: Cooperative IRL

This is a linkpost for https://app.grasple.com/#/level/1606

Every Monday for 8 weeks, we will be posting lessons about Inverse Reinforcement Learning. This is lesson 7.

Note that access to the lessons requires creating an account here.

Have a nice day!

IRL 5/8: Maximum Causal Entropy IRL

This is a linkpost for https://app.grasple.com/#/level/1542

Every Monday for 8 weeks, we will be posting lessons about Inverse Reinforcement Learning. This is lesson 5. We are publishing it now because it would have interfered with the randomized controlled trial we are running, and yesterday we finished collecting responses from the participants. A LW post with the results will appear in a few days. Future IRL lessons will resume normally on Monday.

Note that access to the lessons requires creating an account here.

This lesson comes with the following supplementary material:

- The Principle of Maximum Causal Entropy

Have a nice day!

[AN #54] Boxing a finite-horizon AI system to keep it unambitious

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter.

The newsletter now has exactly 1,000 subscribers! It's a perfect time to **take the 3-minute survey** if you haven't already -- just think of how you'll be making the newsletter better for all 1,000 subscribers! Not to mention the readers on Twitter and the Alignment Forum.

Highlights

Asymptotically Benign AGI *(Michael Cohen)*: I'm a bit out of my depth with this summary, but let's give it a shot anyway. The setting: we are not worried about how much compute we use (except that it should be finite), and we would like to build a powerful AI system that can help us with tasks but does not try to influence the world. We'll assume that we can construct a box from which no signals can pass through, except by pressing a specific button that opens a door.

First, the simple version of BoMAI (Boxed Myopic AI). We'll put the AI system **and the operator** in the box, and the operator and the AI system can talk via text message, and the operator can enter rewards. Each episode has a maximum length (hence myopic), and if the operator ends the episode early, all future rewards are set to zero. BoMAI maximizes episodic reward in a manner similar to AIXI. It has a distribution (initially a speed prior) over all possible time-bounded Turing Machines as possible models that predict observations and rewards. BoMAI uses the maximum a posteriori (MAP) Turing Machine to predict future observations and rewards given actions, searches over all possible sequences of actions for the best one, and executes the first action of that sequence. (See this comment and its parents for the reason to use the MAP model.)

Intuitively, BoMAI has no incentive to affect anything outside the box: information can only leave the box if the episode ends, but if that happens, reward is guaranteed to be exactly zero. It might deceive the operator into giving it high reward, but any such deception will only *incidentally* affect the outside world, it hasn't been *optimized* to cause any such effect. However, **this is all assuming that the MAP world model is accurate**, which may not be true.

Obviously at initialization our world model will not be accurate: we need to explore. Rather than using something like epsilon-exploration, BoMAI delegates to a human explorer for safety, and reduces exploration over time.

More insidiously, a world model could simulate the world for a year, and then report the value in the computer's memory for "the reward from a year ago". This will usually be a perfect prediction of the reward the operator will give, but if BoMAI tries to

optimize for it, it has an incentive to try to e.g. cause the operator to build an unboxed AI whose goal is to change BoMAI's memory cell to the maximum value.

These sorts of bad plans can only happen finitely often. If BoMAI executes such a plan, it will then get the actual reward from the operator, see that it differs from the predicted reward, and update against the world model for making a bad prediction. Eventually the correct model will become more likely. But ideally this sort of thing wouldn't happen at all.

Intuitively, such world models would have to simulate the world past the end of the episode, incurring a speed penalty relative to the correct model. However, they could use that simulation to save on the computation needed for future predictions. To make sure that they do incur a speed penalty, we need to make it so that sometimes their simulations are wrong and can't be reused. So, at the end of each episode, we sample from a source of true randomness, and use that to determine whether or not the operator is given a drug that causes retrograde amnesia. In that case, world models that simulate the future will sometimes be wrong about how that decision plays out, and any simulation they did in the wrong branch is useless and leads to a speed penalty.

With a few non-trivial assumptions, we can prove that asymptotically, BoMAI will do at least as well as the human explorer at accumulating reward, and the MAP world model's rewards do not incentivize BoMAI to affect the outside world.

Rohin's opinion: I think the idea of putting the operator in the box with the AI system is very interesting: with previous attempts at boxing, the human operator talking to the AI system was an obvious glaring hole in the box. In this setting, the only information escaping from the box is the fact that the operator has not yet chosen to end the episode.

I am generally skeptical of intuitive reasoning about what can or can't be done by Turing Machines using extreme amounts of computation. There are *lots* of comments on the post that debate specifics of this. This usually cashes out as a debate about the assumptions in the proof. But it's also worth noting that the theorem is asymptotic, and allows for arbitrarily bad behavior early on. We might still expect good behavior early on for the reasons laid out in the proof, but it's not implied by the theorem, even if the assumptions hold.

Previous newsletters

AI Safety workshop at IJCAI 2019 *(Huáscar Espinoza et al)*: Previously (AN #49), I said the paper submission deadline was April 12. Either I made a mistake, or the deadline has been extended, because the actual deadline is May 12.

Technical AI alignment

Technical agendas and prioritization

AI Alignment Podcast: An Overview of Technical AI Alignment: Part 1 and Part 2 *(Lucas Perry and Rohin Shah)*: In this podcast, I go through a large swath of research agendas around technical AI alignment. The first part is more of a description of what research

agendas exist, who works on them, and what they are trying to do, while the second part delves more into the details of each approach. I'd strongly recommend listening to them if you're trying to orient yourself in the technical AI safety landscape.

Topics covered include embedded agency, value learning, impact regularization methods (AN #49), iterated amplification, debate (AN #5), factored cognition (AN #36), robustness (AN #43), interpretability (no canonical link, but activation atlases (AN #49) is an example), comprehensive AI services (AN #40), norm following (AN #3), and boxing (this newsletter).

Learning human intent

Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations *(Daniel S. Brown, Wonjoon Goo et al)* (summarized by Cody): This paper claims to demonstrate a technique by which an agent learning from a demonstrator's actions can learn to outperform that demonstrator on their true reward, rather than, in the way of imitation learning or behavioral cloning, just mimicking the demonstrator under the assumption that the demonstrator's performance is optimal (or at least near-optimal). The key structural innovation of the paper is to learn using pairs of ranked trajectories and learn a neural network-based reward function based on correctly predicting which will be higher. This allows the model to predict what actions will lead to higher and lower reward, and to extrapolate that relationship beyond the best demonstration. When an agent is then trained using this reward model as it's ground truth reward, it's shown to be capable of outperforming the demonstrator on multiple tested environments, including Atari. An important distinction compared to some prior work is the fact that these rankings are collected in an off-policy manner, distinguishing it from Deep RL from Human Preferences where rankings are requested on trajectories generated as an agent learns.

Cody's opinion: Seems like potentially a straightforward and clever modification to a typical reward learning structure, but a bit unclear how much of the performance relative to GAIL and BCO derives from T-REX's access to suboptimal demonstrations and subtrajectories giving it more effective training data. It does intuitively seem that adding examples of what poor performance looks like, rather than just optimal performance, would add useful informative signal to training. On a personal level, I'm curious if one implication of an approach like this is that it could allow a single set of demonstration trajectories to be used in reward learning of multiple distinct rewards, based on different rankings being assigned to the same trajectory based on the reward the ranker wants to see demonstrated.

Rohin's opinion: It's pretty interesting that the Deep RL from Human Preferences approach works even with off-policy trajectories. It seems like looking at the *difference* between good and bad trajectories gives you more information about the true reward that generalizes better. We saw similar things in our work on Active Inverse Reward Design (AN #24).

End-to-End Robotic Reinforcement Learning without Reward Engineering *(Avi Singh et al)* (summarized by Cody): This paper demonstrates an approach that can learn to perform real world robotics tasks based not on example trajectories (states and actions) but just a small number (10) of pixel-level images of goal states showing successful task completion. Their method learns a GAN-like classifier to predict whether a given image is a success, continually adding data sampled from the still-

learning policy to the set of negative examples, so the model at each step needs to further refine its model of success. The classifier, which is used as the reward signal in learning the policy, also makes use of a simple active learning approach, choosing the state its classifier is most confident is success and querying a human about it on fixed intervals, ultimately using less than 75 queries in all cases.

Cody's opinion: This is a result I find impressive, primarily because of its interest in abiding by sensible real-world constraints: it's easier for humans to label successful end states than to demonstrate a series of actions, and the number of queries made was similarly pragmatically low.

Reward learning theory

AI Alignment Problem: “Human Values” don’t Actually Exist *(avturchin)*

Verification

Optimization + Abstraction: A Synergistic Approach for Analyzing Neural Network Robustness *(Greg Anderson et al)*

AI strategy and policy

Global AI Talent Report 2019 *(Jean-Francois Gagné)*: This report has a lot of statistics on the growth of the field of AI over the last year.

FLI Podcast: Why Ban Lethal Autonomous Weapons? *(Ariel Conn, Emilia Javorsky, Bonnie Docherty, Ray Acheson, and Rasha Abdul Rahim)*

Other progress in AI

Reinforcement learning

How to Train Your OpenAI Five *(OpenAI)*: OpenAI Five (AN #13) has now beaten the Dota world champions 2-0, after training for 8x longer, for a total of 800 petaflop/s-days or 45000 years of Dota self-play experience. During this insanely long training run, OpenAI grew the LSTM to 4096 units, added buybacks to the game, and switched versions twice. Interestingly, they found it hard to add in new heroes: they could bring a few new heroes up to 95th percentile of humans, but it didn't look like they would train fast enough to reach pro level. This could be because the other heroes were already so capable that it was too hard to learn, since the new heroes would constantly be beaten. The resulting team was also able to play cooperatively with humans, even though they had never been trained with humans.

As usual, I like Alex Irpan's thoughts. On the Dota side, he found Five's reaction times more believable, but was disappointed by the limited hero pool. He also predicted that with OpenAI Five Arena, which allowed anyone to play either alongside Five, or against Five, at least one of the *many* teams would figure out a strategy that could reliably beat Five. He was right: while Five had a 99.4% win rate, one team was able to beat it 10 times in a row, another beat it thrice in a row, and two teams beat it twice in a row.

Rohin's opinion: In this era of scaling up compute via parallelism, it was quite surprising to see OpenAI scaling up compute simply by training for almost a year. That feels like one of the last resorts to scale up compute, so maybe we're seeing the limits of the trend identified in AI and Compute (AN #7)?

Back when OpenAI Five beat a strong team in their Benchmark (AN #19), I and a few others predicted that the team would be able to beat Five after playing a few games against it. I think this prediction has been somewhat validated, given that four teams figured out how to beat a much stronger version of the bot. Of course, humans played over 7000 games against Five, not just a few, so this could be that enough random search finds a weakness. Still, I'd expect pros to be able to do this in tens, maybe hundreds of games, and probably this would have been much easier at the time of the Benchmark.

The underlying model here is that Dota has an extremely large space of strategies, and neither Five nor humans have explored it all. However, pros have a better (lower-dimensional) representation of strategy space (concepts like "split-push") that allow them to update quickly when seeing a better opponent. I don't know what it would take to have AI systems learn these sorts of low-dimensional representations, but it seems key to having AI systems that can adapt quickly like humans can.

Read more: Vox: AI triumphs against the world's top pro team in strategy game Dota 2

Deep learning

Do we still need models or just more data and compute? *(Max Welling)*: This is a response to The Bitter Lesson (AN #49), that emphasizes the importance of data in addition to compute. It brings up a number of considerations that seem important to me, and is worth reading if you want to better understand my position on the bitter lesson.

Semantic Image Synthesis with Spatially-Adaptive Normalization *(Taesung Park et al.)* (summarized by Dan H): This paper shows how to create somewhat realistic images specified by semantic segmentation maps. They accomplish this by modifying batch normalization. Batch normalization modifications can be quite powerful for image generation, even enough to control style. Their modification is that normalization is a direct function of the semantic segmentation map throughout the network, so that the semantic segmentation map is readily available to each ResBlock. Visualizations produced by this method are here.

News

SafeML Workshop: Accepted Papers: The camera-ready papers from the SafeML workshop are now available! There are a lot of good papers on robustness, adversarial examples, and more that will likely never make it into this newsletter (there's only so much I can read and summarize), so I encourage you to browse through it yourself.

Why the world's leading AI charity decided to take billions from investors *(Kelsey Piper)*

Read more: OpenAI LP (AN #52)

Many maps, Lightly held

Original post: http://bearlamp.com.au/many-maps-lightly-held/

Many maps, lightly held.

As described in The fox and the hedgehog, among other places (munger, systems thinking). This post holds the theory statement above quite "**strong**", to try to clarify the need for it. It does not apply in some places. For example gravity. It would be difficult to hold gravity lightly although it's a neat thought experiment to wonder how brains and thinking might develop differently in a place that didn't have (almost perfectly) uniform gravity.

I wish I could say the concept of many maps, lightly held was mentioned in Lens that sees it's flaws– but it was not. I believe many maps would fit that post if it were around at the time.

1.

A group of blind men heard that a strange animal, called an elephant, had been brought to the town, but none of them were aware of its shape and form. Out of curiosity, they said: "We must inspect and know it by touch, of which we are capable". So, they sought it out, and when they found it they groped about it. In the case of the first person, whose hand landed on the trunk, said "This being is like a thick snake". For another one whose hand reached its ear, it seemed like a kind of fan. As for another person, whose hand was upon its leg, said, the elephant is a pillar like a tree-trunk. The blind man who placed his hand upon its side said the elephant, "is a wall". Another who felt its tail, described it as a rope. The last felt its tusk, stating the elephant is that which is hard, smooth and like a spear.

If I was a blind man feeling at an elephant, I'd need the principle of many maps to make sense of the world and the information it presented. How can the elephant be both a rope and a spear and a wall? *Many maps. Lightly held.*

2.

When the platypus was first encountered by Europeans in 1798, a pelt and sketch were sent back to Great Britain by Captain John Hunter, the second Governor of New South Wales. British scientists' initial hunch was that the attributes were a hoax. George Shaw, who produced the first description of the animal in the Naturalist's Miscellany in 1799, stated it was impossible not to entertain doubts as to its genuine nature, and Robert Knox believed it might have been produced by some Asian taxidermist. It was thought that somebody had sewn a duck's beak onto the body of a beaver-like animal. Shaw even took a pair of scissors to the dried skin to check for stitches.– Wikipedia page for platypus.

3.

Identity, Archetypes, Roles (mother, teacher, boss). A person can hold many masks in the categories of identities, archetypes or roles. This is an important and valuable feature: to be able to subscribe to a category. The phrase, "*I am a rationalist*", offers a

lot of information. Paul grahamsuggests, “people can never have a fruitful argument about something that’s part of their identity. By definition they’re partisan”.

4.

In philosophical realism, there is a problem between the split of the information that can be found inside the brain and the information outside the brain. If we rely only on information outside the brain, then we are proposing that the information inside the brain is entirely useless. We should collect external information and ignore internal information. This feels like a dangerous trap, there are far too many depressed people to follow external-only reasoning. If we imagine we live in a chinese room, we can’t possibly know if reality is true – through our camera eyeballs and other sensory devices, for all we know we could be living in a simulation. But this doesn’t feel like a complete picture either.

5.

A short experiment in mysticism. Hold your breath. For as long as you can. While you do that, watch your perception of the world. Watch as it gets heavier, denser, feel the redness in the face, feel the tension of the pressure on the chest. Feel the sense of reality closing in. And whichever other perceptions you noticed by testing out this state of experience. Science would happily talk about the (upper right quadrant) phenomena of the body. The carbon dioxide build up, the oxygen depletion, the heart rate change, the body temperature change. Oh science! Beautiful science! I love science. Science is hiding something interesting here behind known maps. Yes, I know the objective maps of what happens when I hold my breath. But do I know the subjective map? What happens to my interior subjective experience when I hold my breath, when I meditate, when I am under stress, when I have an unhealthy diet? How do I know and deal with the subjective without knowing the subjective in great detail? (and I don’t get the knowledge of the subjective from only trying out holding my breath, although it is a neat experiment).

6.

The fable of the rational vampire. (I wish I had a link to credit the author). The rational vampire casually goes through life rationalising away the symptoms – “I’m allergic to garlic”, “I just don’t like the sun”. “It’s impolite to go into someone’s home uninvited, I’d be mortified if I did that”. “I don’t take selfies” and on it goes. Constant rationalisation.

Each of these problems NEEDS many maps. To escape the trap of the flawed lens, I need to be resting in a world of many possible lenses. I need to be willing to hypothesise and entertain that I am a vampire, explaining away my symptoms as if they were allergies and preferences, As well as the concept of being allergic to garlic. The territory only has one explanation but there are many possible maps.

I need to be willing to consider that I am a brain in a box somewhere – and all the signals of the real world are irrelevant. And! Still eat healthy because in the case that I do live in the realism world, I need to be prepared for that too.

I need to be willing to pet the elephant ear, and the elephant trunk and believe it’s one animal if the evidence says so.

I need to live in the world where I am skeptical of the existence of platypuses and willing to check for stitches but also live in a world where it's possible to believe in their existence **at the same time.**

If I want to exist above identities, I need to be willing to be not just my identity, but every other identity too. I need to be able to safely go to the places of uncomfortable identities and wonder why people occupy them. I need to know that I can never take off some of these masks but at least I can know that I am wearing them.

How *many maps, lightly held*, do I use every day…

Scrying for outcomes where the problem of deepfakes has been solved

(Prompted by the post: On Media Synthesis: An Essay on The Next 15 Years of Creative Automation, where Yuli comments "Deepfakes exist as the tip of the warhead that will end our trust-based society")

There *are* answers to the problem of deepfakes. I thought of one, very soon after first hearing about the problem. I later found that David Brin spoke of the same thing 20 years ago in *The Transparent Society*. The idea seems not to have surfaced or propagated at all in any of the deepfake discourse, and I find that a little bit disturbing. There is a cartoon Robin Hanson that sits on my shoulder who's wryly whispering "Fearmongering is not about preparation" and "News is not about informing". I hope it isn't true. Anyway.

In short, if we want to stay sane, we will start building **cameras with tamperproof seals that sign the data they produce with a manufacturer's RSA signature** to verify that the footage comes directly from a real camera, and we will **require all news providers to provide a checked (for artifacts of doctoring and generation), verified, signed (unedited) online copy of any footage they air**. If we want to be extra thorough (and we should), we will also **allocate public funding to the production of disturbing, surreal, inflammatory, but socially mostly harmless deepfakes to exercise the public's epistemic immune system**, ensuring that they remain vigilant enough to check the national library of evidence for signed raws before acting on any new interesting video. I'm sure you'll find many talented directors who'd jump at the chance to produce these vaccinating works, and I think the tradition will find plenty of popular support, if properly implemented. The works could be great entertainment, as will the ensuing identification of dangerously credulous fools.

Technical thoughts about those sealed cameras

The camera's seal should be fragile. When it's broken (*~ when there is any slight shift in gas pressure or membrane conductivity, when the components move, when the unpredictable, randomly chosen build parameters fall out of calibration*), the camera's registered private key will be thoroughly destroyed, with a flash of UV, current, and, ideally, magnesium fire, so that it cannot be extracted and used to produce false signatures. It may be common for these cameras to fail spontaneously. We can live with that. Core components of cameras will mostly only continue to get cheaper.

I wish I could discuss practical processes for ensuring, through auditing, that the cameras' private keys are being kept secret during manufacture. We will need to avoid a situation where manufacturing rights are limited to a small few and the price of authorised sealed cameras climbs up into unaffordable ranges, making them inaccessible to the public and to smaller news agencies, but I don't know enough about the industrial process to discuss that.

(Edit: It occurs to me that the manufacturing process would not have to *inject* a private key from the outside, they would never need to be given access to the private key at all. Sealed camera components can be given a noise generation module to generate their key themselves *after* manufacture is complete. They can then communicate their public key to the factory, and it will be publicly registered as one of the manufacturer's cameras. Video signatures can be verified by finding their pubkey in the manufacturer's registry.)

There's an attack I'm not sure how to address, too. Very high-resolution screens and lenses could be used to show a sealed camera a scene that doesn't exist. The signature attests that the camera genuinely sees it, but it still isn't real. I'll name it the Screen Illusion Analogue Hole Attack (SIAHA).

It might be worth considering putting some kind of GPS chip inside the sealed portion of the camera so that the attack's illusion screen would need to be moved to the location where the fake event was supposed to have happened, which would limit the applications of such an attack, but GPS is currently very easy to fool, so we'll need to find a better location-verification technology than GPS (This is not an isolated need)

I initially imagined that a screen of sufficient fidelity, framerate, and dynamic range would be prohibitively expensive to produce. Now it occurs to me that the VR field aspires to make such screens ubiquitous. [Edit: in retrospect, I think the arguments against commercial VR-based SIAHA I've come up with here are pretty much fatal. It wont happen. There's too much of a difference between what cameras can see and what screens for humans can produce. If a SIAHA screen can be made, it'll be very expensive.]

- Resolution targets may eventually be met.
    - A point in favour: maximum human-perceptible pixel density will be approached. Eye tracking will open the way to foveated rendering; wherein only the small patch of the scene the user is looking directly at will be rendered at max resolution. Current rendering hardware is already beefy enough to support foveated rendering, as it allows us to significantly down-spec the resolution of everything the user isn't looking at. The hardware will not necessarily be made to accept streaming 4k raw footage fresh out of the box (more like 720p footage and another patch of 720p footage for the foveal patch), but the pixels will all be there, the screen will be dense enough, it will be very possible to produce hardware that will do it, if not by modifying a headset, then by modifying its factory.
    - A point against: Video cameras will sometimes want to go beyond retinal pixel density for post-production digital zoom. They will want to capture much more than a human standing in their position can see, and I can see no reason consumer-grade screens should ever come to output more detail than a human can see.
- Framerate targets will be met because if you dip below 100fps in VR, players puke. It's a hard requirement. There will never be a commercial VR headset that couldn't do it.
    - point against: If the framerate of the screen is not much much higher than the framerate of the camera, if they're merely similar, unless they're perfectly synchronised, frame skips or tears will occur.
- Realistic dynamic range might take longer than the other two, but there will be a demand for it... though perhaps we will never want a screen that can flash with the brightness of the sun. If cameras of the future can record that level of

brightness, that may be some defence against this kind of attack, at least for outdoor scenes.
- Color accuracy may remain difficult to replicate with screens. Cameras already accidentally record infra red light. Screens for humans will never need to produce infra-red. I'm not sure how current cameras' color accuracy compares to the human eye... I suspect it's higher, but I'm not able to confirm that.

[in conclusion, I think cheap SIAHA is unlikely]

In summary: A combination of technologies, laws, and fun social practices can probably mostly safeguard us against the problem of convincingly doctored video evidence. Some of the policing and economic challenges are a bit daunting, but not obviously insoluble.

Alignment Newsletter #52

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter.

Highlights

Thoughts on Human Models *(Ramana Kumar and Scott Garrabrant)*: Many approaches to AI safety involve modeling humans in some way, for example in order to correctly interpret their feedback. However, there are significant disadvantages to human modeling. First and most importantly, if we have AI systems do useful things *without* modeling humans, then we can use human approval as a "test set": we can check whether the AI's behavior is something we approve of, and this is an *independent* evaluation of the AI system. However, if the AI system had a human model, then it may have optimized its behavior for human approval, and so we cannot use approval as a "test set". Second, if our AI system has a catastrophic bug, it seems better if it doesn't have any human models. An AI system without human models will at worst optimize for some unrelated goal like paperclips, which at worst leads to it treating humans as obstacles and causing extinction. However, an AI system with human models with a catastrophic bug might optimize for human suffering, or having humans respond to email all day, etc. Thirdly, an AI system with human models might be simulating conscious beings that can suffer. Fourthly, since humans are agent-like, an AI system that models humans is likely to produce a subsystem that is agent-like and so dangerous.

The authors then discuss why it might be hard to avoid human models. Most notably, it is hard to see how to use a powerful AI system that avoids human models to produce a better future. In particular, human models could be particularly useful for interpreting specifications (in order to do what humans mean, as opposed to what we literally say) and for achieving performance given a specification (e.g. if we want to replicate aspects of human cognition). Another issue is that it is hard to avoid human modeling, since even "independent" tasks have some amount of information about human motivations in selecting that task.

Nevertheless, the authors would like to see more work on engineering-focused approaches to AI safety without human models, especially since this area is neglected, with very little such work currently. While MIRI does work on AI safety without human models, this is from a very theoretical perspective. In addition to technical work, we could also promote certain types of AI research that is less likely to develop human models "by default" (e.g. training AI systems in procedurally generated simulations, rather than on human-generated text and images).

Rohin's opinion: While I don't disagree with the reasoning, I disagree with the main thrust of this post. I wrote a long comment about it; the TL;DR is that since humans want very specific behavior out of AI systems, the AI system needs to get a lot of information from humans about what it should do, and if it understands all that information then it necessarily has a (maybe implicit) human model. In other words, if

you require your AI system not to have human models, it will not be very useful, and people will use other techniques.

Technical AI alignment

Iterated amplification

AI Alignment Podcast: AI Alignment through Debate *(Lucas Perry and Geoffrey Irving)* (summarized by Richard): We want AI safety solutions to scale to very intelligent agents; debate is one scalability technique. It's formulated as a two player zero-sum perfect information game in which agents make arguments in natural language, to be evaluated by a human judge. Whether or not such debates are truth-conducive is an empirical question which we can try to evaluate experimentally; doing so will require both technical and social science expertise (as discussed in a previous post (AN #47)).

Richard's opinion: I think one of the key questions underlying Debate is how efficiently natural language can summarise reasoning about properties of the world. This question is subject to some disagreement (at one extreme, Facebook's roadmap towards machine intelligence describes a training environment which is "entirely linguistically defined") and probably deserves more public discussion in the context of safety.

Rohin's note: If you've read the previous posts on debate, the novel parts of this podcast are on the relation between iterated amplification and debate (which has been discussed before, but not in as much depth), and the reasons for optimism and pessimism about debate.

Agent foundations

Pavlov Generalizes *(Abram Demski)*: In the iterated prisoner's dilemma, the Pavlov strategy is to start by cooperating, and then switch the action you take whenever the opponent defects. This can be generalized to arbitrary games. Roughly, an agent is "discontent" by default and chooses actions randomly. It can become "content" if it gets a high payoff, in which case it continues to choose whatever action it previously chose as long as the payoffs remain consistently high. This generalization achieves Pareto optimality in the limit, though with a very bad convergence rate. Basically, all of the agents start out discontent and do a lot of exploration, and as long as any one agent is discontent the payoffs will be inconsistent and all agents will tend to be discontent. Only when by chance all of the agents take actions that lead to all of them getting high payoffs do they all become content, at which point they keep choosing the same action and stay in the equilibrium.

Despite the bad convergence, the cool thing about the Pavlov generalization is that it only requires agents to notice when the results are good or bad for them. In contrast, typical strategies that aim to mimic Tit-for-Tat require the agent to reason about the beliefs and utility functions of other agents, which can be quite difficult to do. By just focusing on whether things are going well for themselves, Pavlov agents can get a lot of properties in environments with other agents that Tit-for-Tat strategies don't obviously get, such as exploiting agents that always cooperate. However, when thinking about logical time (AN #25), it would seem that a Pavlov-esque strategy would have to make decisions based on a prediction about its own behavior, which

is... not obviously doomed, but seems odd. Regardless, given the lack of work on Pavlov strategies, it's worth trying to generalize them further.

Approval-directed agency and the decision theory of Newcomb-like problems *(Caspar Oesterheld)*

Learning human intent

Thoughts on Human Models *(Ramana Kumar and Scott Garrabrant)*: Summarized in the highlights!

Verification

Algorithms for Verifying Deep Neural Networks *(Changliu Liu et al)*: This is a survey paper about verification of properties of deep neural nets.

Robustness

Towards Robust and Verified AI: Specification Testing, Robust Training, and Formal Verification *(Pushmeet Kohli et al)*: This post highlights three areas of current research towards making robust AI systems. First, we need better evaluation metrics: rather than just evaluating RL systems on the environments they were trained on, we need to actively search for situations in which they fail. Second, given a specification or constraint that we would like to ensure, we can develop new training techniques that can ensure that the specifications hold. Finally, given a specification, we can use formal verification techniques to ensure that the model obeys the specification on all possible inputs. The authors also list four areas of future research that they are excited about: leveraging AI capabilities for evaluation and verification, developing publicly available tools for evaluation and verification, broadening the scope of adversarial examples beyond the L-infinity norm ball, and learning specifications.

Rohin's opinion: The biggest challenge I see with this area of research, at least in its application to powerful and general AI systems, is how you get the specification in the first place, so I'm glad to see "learning specifications" as one of the areas of interest.

If I take the view from this post, it seems to me that techniques like domain randomization, and more generally training on a larger distribution of data, would count as an example of the second type of research: it is a change to the training procedure that allows us to meet the specification "the agent should achieve high reward in a broad variety of environments". Of course, this doesn't give us any provable guarantees, so I'm not sure if the authors of the post would include it in this category.

Forecasting

Historical economic growth trends *(Katja Grace)* (summarized by Richard): Data on historical economic growth "suggest that (proportional) rates of economic and population growth increase roughly linearly with the size of the world economy and population", at least from around 0 CE to 1950. However, this trend has not held since 1950 - in fact, growth rates have fallen since then.

Miscellaneous (Alignment)

Coherent behaviour in the real world is an incoherent concept *(Richard Ngo)*: In a previous post (AN #35), I argued that coherence arguments (such as those based on VNM rationality) do not constrain the behavior of an intelligent agent. In this post, Richard delves further into the argument, and considers other ways that we could draw implications from coherence arguments.

I modeled the agent as having preferences over full trajectories, and objected that if you only look at *observed* behavior (rather than *hypothetical* behavior), you can always construct a utility function such that the observed behavior optimizes that utility function. Richard agrees that this objection is strong, but looks at another case: when the agent has preferences over states at a single point in time. This case leads to other objections. First, many reasonable preferences cannot be modeled via a reward function over states, such as the preference to sing a great song perfectly. Second, in the real world you are never in the same state more than once, since at the very least your memories will change, and so you can never infer a coherence violation by looking at observed behavior.

He also identifies further problems with applying coherence arguments to realistic agents. First, all behavior is optimal for the constant zero reward function. Second, any real agent will not have full information about the world, and will have to have beliefs over the world. Any definition of coherence will have to allow for multiple beliefs -- but if you allow all beliefs, then you can rationalize any behavior as based on some weird belief that the agent has. If you require the agent to be Bayesian, you can still rationalize any behavior by choosing a prior appropriately.

Rohin's opinion: I reject modeling agents as having preferences over states primarily for the first reason that Richard identified: there are many "reasonable" preferences that cannot be modeled with a reward function solely on states. However, I don't find the argument about beliefs as a free variable very convincing: I think it's reasonable to argue that a superintelligent AI system will on average have much better beliefs than us, and so anything that we could determine as a coherence violation with high confidence should be something the AI system can also determine as a coherence violation with high confidence.

Three ways that "Sufficiently optimized agents appear coherent" can be false *(Wei Dai)*: This post talks about three ways that agents could not appear coherent, where here "coherent" means "optimizing for a *reasonable* goal". First, if due to distributional shift the agent is put into situations it has never encountered before, it may not act coherently. Second, we may want to "force" the agent to pretend as though compute is very expensive, even if this is not the case, in order to keep them bounded. Finally, we may explicitly try to keep the agent incoherent -- for example, population ethics has impossibility results that show that any coherent agent must bite some bullet that we don't want to bite, and so we may instead elect to keep the agent incoherent instead. (See Impossibility and Uncertainty Theorems in AI Value Alignment (AN #45).)

The Unavoidable Problem of Self-Improvement in AI and The Problem of Self-Referential Reasoning in Self-Improving AI *(Jolene Creighton and Ramana Kumar)*: These articles introduce the thinking around AI self-improvement, and the problem of how to ensure that future, more intelligent versions of an AI system are just as safe as the original system. This cannot be easily done in the case of proof-based systems, due to Godel's incompleteness theorem. Some existing work on the problem: Botworld, Vingean reflection, and Logical induction.

Other progress in AI

Deep learning

The Lottery Ticket Hypothesis at Scale *(Jonathan Frankle et al)* (summarized by Richard): The lottery ticket hypothesis is the claim that "dense, randomly-initialized, feed-forward networks contain subnetworks (winning tickets) that - when trained in isolation - reach test accuracy comparable to the original network in a similar number of iterations". This paper builds on previous work to show that winning tickets can also be found for larger networks (Resnet-50, not just Resnet-18), if those winning tickets are initialised not with their initial weights from the full network, but rather with their weights after a small amount of full-network training.

Richard's opinion: It's interesting that the lottery ticket hypothesis scales; however, this paper seems quite incremental overall.

News

OpenAI LP *(OpenAI)* (summarized by Richard): OpenAI is transitioning to a new structure, consisting of a capped-profit company (OpenAI LP) controlled by the original OpenAI nonprofit organisation. The nonprofit is still dedicated to its charter, which OpenAI LP has a legal duty to prioritise. All investors must agree that generating profits for them is a secondary goal, and that their overall returns will be capped at 100x their investment (with any excess going back to the nonprofit).

Richard's opinion: Given the high cost of salaries and compute for machine learning research, I don't find this a particularly surprising development. I'd also note that, in the context of investing in a startup, a 100x return over a timeframe of decades is not actually that high.

Quantitative Philosophy: Why Simulate Ideas Numerically?

This is a linkpost for https://quantitativephilosophy.wordpress.com/

Adapted from my blog. I argue that numerical simulations are an effective yet underused tool in philosophy (and rationality) and give a concrete example of a numerical toy model, from assumptions to design to implementations to results, including the source code and a live site to play with.

It used to be that logic was king. A convincing argument was all that was necessary to get one's ideas taken seriously and often accepted. After all, if it makes sense, it must be right, right? Right? The thought of actually checking the ideas experimentally is not very old, and, while firmly entrenched in the scientific method, sometimes earnestly, sometimes as a lip service, it is still not the commonly accepted practice in many "softer" sciences.

This forum is a site devoted to rationality, and has plenty of interesting insights, but very few of them have been actually tested. There is a good reason for it: thinking ideas up is much easier than checking them! Eliezer Yudkowsky, the original contributor, set this tone of using reasoning as an argument, without ever closing the feedback loop of checking the conclusions with numbers. There were a few exceptions to this pattern, such as the Iterated Prisoner's Dilemma bot tournaments, but in general the focus is on reasoning and sometimes mathematical proofs, where possible and warranted, not on making testable predictions and actually testing them.

This leaves out one of the most powerful tools of checking validity of an idea: numerical simulation. If the idea is any good, one ought to be able to formalize it to the degree where its conclusions can be tested by creating simulations and studying their behavior.

The situation is somewhat better at Slate Star Codex: Scott Alexander has lots of wonderful ideas, and he is more aware of the need to check them by something other than more logic. But his approach for testing them is generally literature search or polls/surveys. Those are useful, but they focus on the ideas as black boxes, rather than on their internal mechanics.

Some examples of interesting ideas ripe for numerical modeling:

- Inadequate Equilibria
- Motte and Bailey
- Moloch
- Outgroup
- Evaporative Cooling of Group Beliefs

In the following I focus on one idea, similar to the outgroup one listed above:

PEOPLE FEEL MORE AFFINITY TO THOSE WHOSE VIEWS ARE CLOSE TO THEIR OWN, AND ARE OFTEN REPULSED BY THOSE WHOSE VIEWS DIVERGE A LOT FROM THEIRS. WHAT IS THE RESULTING DYNAMICS OF THESE HUMAN PREFERENCES?

An obvious conclusion that people form cohesive groups that are hostile to other cohesive groups. A set of local (often suboptimal in some sense) equilibria forms that is hard to change.

How would one go about modeling these ideas numerically? My attempt at doing so is described below.

First, obviously we cannot model this segmentation and alienation process in full generality, there are too many factors to consider. The art of modeling is in picking those that are both important and easy to describe quantitatively. The results of such an approach look pretty natural in retrospect, but are anything but easy to converge on. So, below I will endeavor to describe not just the final model, but how I got there, without the benefit of a hindsight.

- How does interaction between people work? At a very basic level people exchange ideas, opinions, thoughts and, more often than not, praise and insults.
- This used to happen mostly between those who are in physical proximity, but, the world being global and connected, at least (mis)informationally, the interaction is no longer limited by distance
- Even people who are fairly close in views to each other tend to have a bit of variation, and their views may shift randomly a little here and there

To describe what happens to people's views as a result of an interaction between them, we need some simplified description that can be modeled numerically. People tend to fall prey to the confirmation bias a lot, which gives us two basics characteristics of the interaction:

- Interaction between two people who are fairly close in their views on a given topic provides the confirmation they crave, and results in further convergence of the views. This phenomenon can be described as a force of attraction.
- Interaction between people with wildly divergent views also provides the confirmation of their own opinions, only in contrast to the other person's horribly misguided and wrong one. This can be modeled as a, well, repulsive force, pushing people further apart.

What happens when people's views are very far from or very close to each other?

- When the views have almost converged, there is very little change as a result of the interaction, so the force of attraction is, counter-intuitively, gets weaker at "shorter distances"
- When the views are so far apart, we can no longer relate to them, or maybe even take them seriously, the result of the interaction is very little change in our own views, so the force of repulsion is weaker, as well.
- Potentially, since we are all humans, there are limits to how far the views can diverge, because there are generally some basic shared ideas that most people subscribe to, like, say, the desire for humanity to survive and prosper. So, at the very large divergences the repulsion again turns into attraction, if very weak.

Humans have opinions on a wide variety of topics, and, while there is a correlation, even if they are in agreement on one issue, they do not necessarily agree on every issue. This complicates the situation, making it multi-dimensional, and adding complications to a potential numerical model of it. So, at this point, as a toy model, focusing on a single dimension should be a good enough start.

It is a well known phenomenon that people sometimes radically change their views, for example undergoing religious conversion or de-conversion, and changing group allegiances as a result. It is not immediately clear how to model this numerically, and, unless this process magically emerges from the results of the simulations, is best left for future investigations.

Another point to keep in mind is the initial distribution of views before the interaction. If everyone starts agreeing with everyone else, it is not likely that anything would change, given the above assumptions. So, some spread of the views is necessary to get a non-trivial dynamics. The nature of this spread is probably something to play with once the model is implemented. Potential options to consider:

- Symmetric vs. asymmetric distributions, corresponding to mainstream vs niche views.
- Uniform vs. “dumbbell” distribution, where people are already primed even before the interaction.

Assuming the above program is implemented, what do we expect to learn from running the simulations?

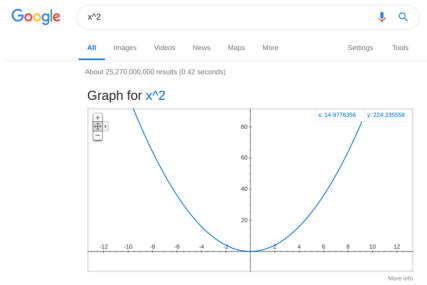
- First, validate the model itself! Always an essential step. If the underlying code or math is wrong, the model is worse than useless, it can force one to make completely unwarranted conclusions. (Totally not speaking from experience! Not at all!)
- Second, figure out the range of parameters where we get the expected results. Every model has adjustable numerical parameters, and finding the parameter space as the home base to start at is definitely worthwhile.
- It might happen that there is no such home base, and this would be even more exciting than getting the expected results! It would either mean that the whole idea is bogus, or that our model of it is inadequate. In either case, it is back to the drawing board!
- Once we have a well-behaving numerical model, it is play time! Time to reap the rewards of all the hard work by varying the parameter space and watching what happens. Hopefully something new and unexpected would show up!
- This is where the real payoff of numerical modeling is: finding something new and having those “Aha!” moments, where the unexpected results make us learn something new and gain insights that were missed in simply “applying logic”.

So having put some thought into what to model, what simplification to make and what the goals are, it’s time to get down and dirty.

I have already decided that:

- The initial version would be one-dimensional.
- The interaction is attractive at short distances, repulsive at large distances, and again weakly attractive at large distances.

These two assumptions let us pick the shape of the interaction. Given my background in physics, I naturally think of forces in terms of potentials first. The magnitude of the force corresponds to the slope of the potential. Attractive force corresponds to the positive slope, and repulsive force corresponds to the negative slope. The complete potential can be combined from those with the basic tools of addition and multiplication. To simulate the attractive force, we can use everyone’s favorite harmonic oscillator: a stretched spring tends to contract. Google helpfully constructs the graph for it:

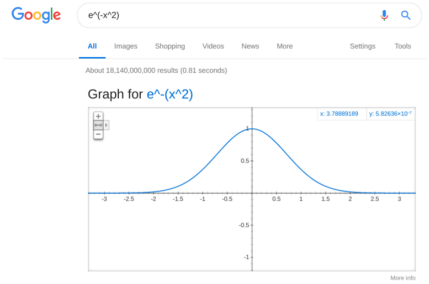


This has the basic shape we want:

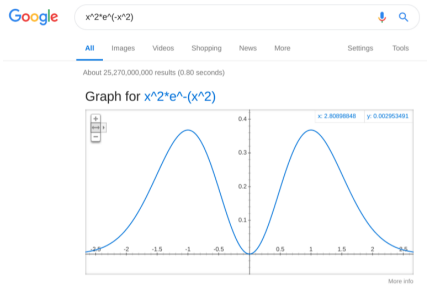
- The attraction (convergence of views) is low when the views are already similar.
- The attraction increases when the separation is a bit larger.

If you are not familiar with the potential curves, think of it as a hilly slope, and you and your friend being on the opposite sides of it. In the situation above you would naturally roll down toward each other.

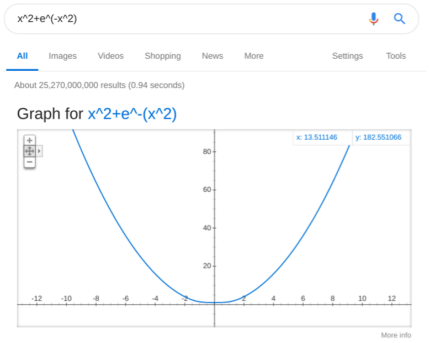
Now, to get the repulsive side of the interaction, we can use another well-worn tool, the bell curve:



So we have the two out of three features of the interaction shape: the attraction when close together and repulsion when further apart. Let’s compose them together, and the way to do it is simple multiplication:



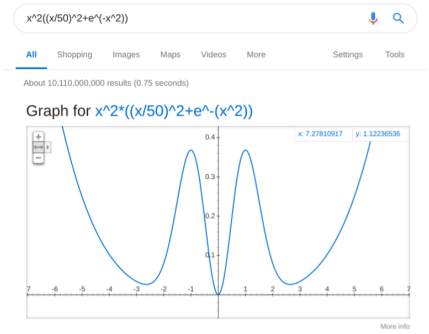
We are almost there! Just need to add the last piece, the weak attraction at large separations. Again, the harmonic oscillator potential to the rescue! Only we need to add it to the bell curve first:



Uh oh… That didn’t work out as expected! The spring is too stiff and completely overwhelms the little bell. Let’s loosen the spring up a bit, by dividing its potential by a big number. Say, make 5 times looser:

That looks better, but clearly the stiffness of the spring should become one of those adjustable parameters once the model is done. Now to put all three together:

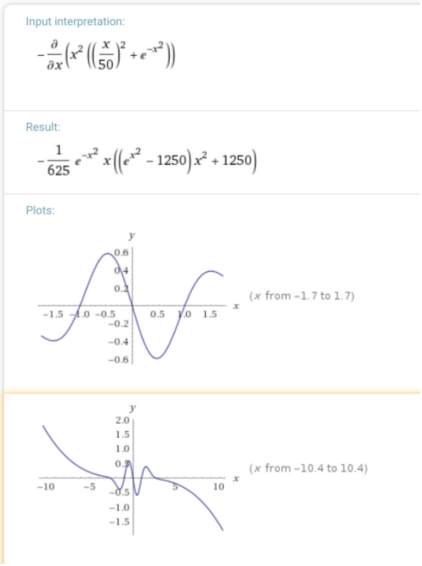
Another oops… What happened? The bell curve, again, was too weak to show prominently. Need to suppress the weaker spring a bit more, maybe 10 times more:



This is more like it! Let's look at it more closely:

- People whose opinions are separated by approximately less than one "unit of disagreement" would tend to "slide" toward each other and maybe slosh a bit around zero-disagreement.
- People whose opinions a separated by somewhat more than that, will find each other's views repulsive enough to feel like the other person is a part of an "outgroup", and instinctively distance themselves from them.
- Eventually the "shared human values" start to matter, and there is a certain separation distance where the two parties, while repulsed by each other, just shrug it off without any further need to distance themselves from the other. How faithful this model is is debatable, of course, but for a first approximation it does not look too out of place.

Clearly playing with the shape of the interaction potential is not done, but it is good enough for now, and it is time to go to the next step. Well, not quite. Let's have one more chart, converting the shape of the potential into the shape of the force. Google calculator is not up to snuff there quite yet, but Wolfram Alpha is, though the free version has low resolution and no customization:



I don't find the force graphs as illuminating as the potential graphs, but one can still make sense of it: negative values correspond to the force pulling to the left, and positive values correspond to the force pulling to the right. So, again, when the views are close, they converge, when they are farther apart, they tend to separate to a

respectable distance, but not infinitely far. The force, and not the potential is what we need to calculate how people’s opinions move after interacting, anyway.

That was the easy part, picking the shape, next we have to figure out how to actually implement the dynamics of interactions and changing personal views.

I will not discuss it here, an interested party can read about it in my blog post:

https://quantitativephilosophy.wordpress.com/2019/04/09/separation-and-clustering-of-views-making-a-step/

The code itself can be found in https://quantitativephilosophy.wordpress.com/2019/04/12/separation-and-clustering-of-views-the-app/

This site doesn't seem to allow embedded sites or embedded html+javascript, so I cannot insert the actual app in here, instead you can go to

https://sites.google.com/view/numericalsimulationofclusterin/home

to play with the model. I will talk about it more in a companion post to follow. But here is a video of one simulation run: https://i.imgur.com/iDaH9nj.mp4 to give the idea of it, wish I could embed it here though.