Best of LessWrong: September 2017

1. For signaling? (Part I)
2. Splitting Decision Theories
3. Common vs Expert Jargon
4. Notes From an Apocalypse
5. Against EA PR
6. Strategic Goal Pursuit and Daily Schedules
7. The Virtue of Numbering ALL your Equations
8. The Outside View isn't magic
9. Sabbath hard and go home
10. The Iron Law of Evaluation and Other Metallic Rules.
11. Stupid Questions - September 2017
12. Exposition and guidance by analogy
13. Impression track records
14. Out to Get You
15. Extensive and Reflexive Personhood Definition
16. Beta - First Impressions
17. The Anthropic Principle: Five Short Examples
18. Against Individual IQ Worries
19. Why I am not a Quaker (even though it often seems as though I should be)
20. Metamathematics and Probability
21. Wikipedia pageviews: still in decline
22. Tests make creating AI hard
23. An incentive structure that might not suck too much.
24. I Can Tolerate Anything Except The Outgroup
25. Happy Petrov Day! If Today is Also Your Birthday, Happy Birthday!
26. Nobody does the thing that they are supposedly doing
27. In Defense of Unreliability
28. Intellectual Progress Inside and Outside Academia
29. Cognitive Empathy and Emotional Labor
30. Post Fit Review
31. Why I Quit Social Media
32. Value Arbitrage
33. Moderator's Dilemma: The Risks of Partial Intervention
34. Frontpage Posting and Commenting Guidelines
35. The Best Self-Help Should Be Self-Defeating
36. Blind Goaltenders: Unproductive Disagreements
37. The Great Filter isn't magic either
38. Motivating a Semantics of Logical Counterfactuals
39. Why Attitudes Matter
40. Musings on Double Crux (and "Productive Disagreement")
41. Epistemic Spot Check: Exercise for Mood and Anxiety (Michael W. Otto, Jasper A.J. Smits)
42. Thinking on the page
43. The Five Hindrances to Doing the Thing
44. Unfair outcomes from fair tests
45. Solomonoff Induction explained via dialog.
46. Map of the AI Safety Community

Best of LessWrong: September 2017

1. For signaling? (Part I)
2. Splitting Decision Theories
3. Common vs Expert Jargon
4. Notes From an Apocalypse
5. Against EA PR
6. Strategic Goal Pursuit and Daily Schedules
7. The Virtue of Numbering ALL your Equations
8. The Outside View isn't magic
9. Sabbath hard and go home
10. The Iron Law of Evaluation and Other Metallic Rules.
11. Stupid Questions - September 2017
12. Exposition and guidance by analogy
13. Impression track records
14. Out to Get You
15. Extensive and Reflexive Personhood Definition
16. Beta - First Impressions
17. The Anthropic Principle: Five Short Examples
18. Against Individual IQ Worries
19. Why I am not a Quaker (even though it often seems as though I should be)
20. Metamathematics and Probability
21. Wikipedia pageviews: still in decline
22. Tests make creating AI hard
23. An incentive structure that might not suck too much.
24. I Can Tolerate Anything Except The Outgroup
25. Happy Petrov Day! If Today is Also Your Birthday, Happy Birthday!
26. Nobody does the thing that they are supposedly doing
27. In Defense of Unreliability
28. Intellectual Progress Inside and Outside Academia
29. Cognitive Empathy and Emotional Labor
30. Post Fit Review
31. Why I Quit Social Media
32. Value Arbitrage
33. Moderator's Dilemma: The Risks of Partial Intervention
34. Frontpage Posting and Commenting Guidelines
35. The Best Self-Help Should Be Self-Defeating
36. Blind Goaltenders: Unproductive Disagreements
37. The Great Filter isn't magic either
38. Motivating a Semantics of Logical Counterfactuals
39. Why Attitudes Matter
40. Musings on Double Crux (and "Productive Disagreement")
41. Epistemic Spot Check: Exercise for Mood and Anxiety (Michael W. Otto, Jasper A.J. Smits)
42. Thinking on the page
43. The Five Hindrances to Doing the Thing
44. Unfair outcomes from fair tests
45. Solomonoff Induction explained via dialog.
46. Map of the AI Safety Community

For signaling? (Part I)

Your T-shirt is embarrassing. Have you considered wearing a less embarrassing T-shirt?

You are suggesting I spend my precious time trying to look good. Well I am good, and so I'm not going to do that. Because signaling is bad. You can tell something is bad when the whole point of it is to have costs. Signaling is showing off. Signaling benefits me at someone else's equal expense. I won't wear a less embarrassing T-shirt because to Hell with signaling.

Hmm. That seems wrong. Signaling is about honest communication when the stakes are high—which is often important! And just because it's called 'costly' doesn't mean it is meant to have costs. It only has to be too costly for liars, and if it's working then they won't be doing any signaling anyway. 'Costly signals' can be very cheap for those who use them. I think signaling is often wonderful for society.

Give me three examples where it is 'wonderful'.

Driver's licences. Showing a driver's licence is a costly signal of being a decent driver, which communicates something useful honestly, is cheap for the people who are actually good drivers, and lets the rest of society distinguish people who are likely to drive safely from people who are not, which is amazingly great.

Driving tests don't seem that cheap to me, but I'll grant that they are probably worth it. Still, this seems like a strange corner case of 'signaling' that was explicitly designed by humans. It fits the economic definition of 'costly signaling' but if you have to go that far from the central examples to find something socially beneficial, that doesn't increase my regard for signaling. Next?

One of the most famous examples of signaling is in the job market. Potential candidates show a hirer their qualifications, which allows the hirer to employ more appropriate candidates. You might disagree about whether all of the signals that people use are socially optimal—for instance if education is mostly for signaling, it seems fairly destructive, because it is so expensive. But you must agree that companies do a lot better hiring the people they choose than they would hiring random people they would get if good candidates couldn't signal their quality. And at least many aspects of the interview process are cheap enough to be totally worth it. For instance, being able to have a polite and friendly conversation about the subject matter.

Of course companies are better off—companies aren't the people destroying years of their productive lives on deliberately arduous fake work. Or learning a lot of irrelevant but testable skills. Or degrading themselves and society with faux friendliness. And you ignore some other key details, like what the actual alternative would realistically look like. But let's not go into it—I'll grant you that hiring probably goes better overall than it would with zero signaling and no replacement, even though the signaling is awful. And more importantly, that the the whole of society on net is probably best off with some kind of signaling there. I don't know of a good replacement.

Ok, great. So, third—T-shirts. T-shirts signal personality traits. It is free to wear any T-shirt you want, but T-shirts are still costly signals in a sense, because if you aren't a punk you won't know which T-shirt to wear to look like a genuine punk. And if you don't like ABBA it is more costly for you to wear an ABBA t-shirt than it is for someone

who does like them, because you'll be embarrassed or unhappy at the association. And if you have bad taste, it is hard to know which T-shirt would indicate good taste. This all seems good, because it lets people cheaply find other people with similar interests, and also to learn facts about the people around them, regardless of similarity. Which is why it is socially destructive for you to wear that T-shirt— your taste can't be that bad, so you are basically lying.

...

Ok, a fourth: how about when a friend is sick, and you make them tea and soup and put on a movie for them. This is a costly signal that you care about them, or at least about your continuing friendship with them. Because it is effort for you with no reward if you don't care much, and are looking to scale down the relationship soon. But aside from the signaling, this is probably a net social benefit—your friend gets soup and tea and a movie at a time when they could especially use them. Plus, feeling cared for instead of uncared for is a real benefit.

Ok, I concede that costly signaling can be honest, cheap, and on net socially beneficial. But I still think it usually isn't! And I'm not sure how far we can get thinking about specific examples, since there are so many.

Ok, what do you propose?

Talking about our overall impressions. The big picture. Here is mine: the world is full of people pouring real wealth into things whose only use is to be rubbed in the face of those who can't afford to destroy so much value. Where it isn't even good for society to be able to distinguish the signalers from the rest. Letting everyone see who is rich and who is poor, who is socially competent and who is not, who is beautiful, who is smart, who can win at things that only exist to be won at—does this really lead to a great world?

There is much signaling that the world would be better off without. I admit I don't really know what the balance of good and bad is like. But I disagree that we should be talking about signaling overall. Or even what is best for the world in this particular case. You are not the world. Even signaling that is terrible for the world is often good for you. If you are in a zero-sum game, and you are more worthy than the opponent, then do your best to win! And if you aren't, then be more worthy!

What if I want what is best for society?

Even then, you don't serve society by failing at signaling. Just because people fighting to look good is costly for society doesn't mean that society gains anything by you intentionally losing that fight. If you are directing your resources to society, then it is better for society if you win. Often better enough to warrant the costs of playing. Serve society by winning at signaling and donating the proceeds to society. Wear a well ironed suit. Don't talk about your erotic porcelain dinosaur collection. Go to university. Try to exercise good taste...

I agree, at least often. But I think you believe in a heuristic that says you should signal about as much and in similar ways as if you were selfish. Because you are on the side of good, so protecting yourself is protecting the good. You see people looking weird and embarrassing themselves in the name of caring about something, and you think they are failing at signaling. And that's wrong.

Yeah, I guess you should signal a tiny bit less on the margin, in cases where signaling is socially destructive. But it's such a small thing, I'm not sure it is worth thinking about.

I don't mean that. Your selfish interests can come apart from society's interests almost entirely, in signaling. As an extreme case, imagine that you became confident that by far the best cause for improving the world was promoting incest. From a selfish perspective, you probably don't want to look like you are promoting incest, because there are few worse ways to look in modern society. But from an altruistic perspective, supposing that you were right about incest, it may well be best for you to promote it, because it would do so much for making incest look better, at just the cost of your own reputation.

You should distinguish between wearing a clean shirt—good for your cause—and wearing a shirt that is more respectable because it is not about your cause—which is often bad for your cause. You can't just use 'looking good' as a heuristic, even though it is generally good for your cause when its proponents look good.

That's an interesting point, and I hadn't really thought about it. But surely that's pretty rare. There are systematic reasons that it's unlikely that there is some cause which is radically more important than any other, and is completely politically unpalateable.

I agree that's unlikely—just brought it up as a clear example of it being not worth looking good. I think this issue is maybe ubiquitous though, in less clear and extreme cases. For instance, everywhere sophisticated people play it cool, withholding enthusiasm from ideas until they no longer lack enthusiasm, polishing their own image at the expense of the very projects they are most excited about, or would be if they deigned to experience excitement.

A bold claim—I am curious to hear two more examples, but I have a lot of signaling to get done this evening. Same time next week?

Most likely. I hope you are correctly identified as the superior type in all of your endeavors.

Splitting Decision Theories

[epistemic status: maybe wrong; thinking aloud, would like people to yell at me]

There is a repeated motion that occurs when deciding what an AI should do:

(1) Create a decision theory

(2) Create a thought experiment in which an agent with *DT makes a choice which fails to fulfill its utility function (e.g. Oh no! It loses all its money to blackmail!)

(3) Create a DT which does well against problems which the core difficulty which allowed the previous decision theory to lose all its money

If decision theories are as precisely imagined as mathematical structures. For every two distinct decision theories, there exists in mathematical reality a set of "thought experiments" such that the two theories decide differently on them.

This seems weird and difficult now because there isn't a shared logical notation between different "thought experiments". As of now characterizing the class of splitting decision problems for two decision theories is pretheoretic. However, for every pair of decision theories DT_1 and DT_2 the object split(DT_1, DT_2) actually exists. Current notational limits make it currently difficult to simply and completely characterize the class of choice problems on which two DTs give different answers.

But it feels like this sort of problem occupies a similar status as “algorithms” did before the first Universal Turing Machine was constructed.

-

Questions:

In fun games (like prisoner’s dilemma) we have agents (like fairbot) that fight each other. The source code for these agents is entangled with their decision theory. Does examining bots engaged in modal combat make this problem more tractable?

This process repeats like clockwork (it feels like a new decision theory comes out every year or so?) in hopes of giving their baby AI a good way of making good choices and not losing all its money. What if I built an AI that formalized and internalized this process and just ... gave itself good advice? Within logical inductors traders bet on which theorems would be best and traders which make bad bets lose their money. If we can formalize split(DT_1, DT_2) we can look at how well agents fulfill their utility functions in this space. Can we use this to establish a kind of poset of decision theories?

Common vs Expert Jargon

tldr: Jargon always has a complexity cost, but you can put effort into making a concept more accessible, and it's especially valuable to put that effort in for terms that you'd like to be used by layfolk, or that you expect to be used a lot in spaces where you expect lots of layfolk to be reading/participating.

I. Lessons from Game Design

Magic the Gathering deals a lot with complexity. Each year, new abilities and rules are added to the game. This gives experienced players the chance to constantly discover new things, but it comes with some issues.

First, it makes the game harder for new players (the game kept growing more complex over time, raising the amount of information a new player had to process at once)

And second, even for experienced players: each instance of complexity is a cost. Players (both new and old) can only handle so much, and some forms of complexity are less fun than others. (For example, forcing players to do a lot of book-keeping, rather than letting them make interesting strategic decisions)

Six years ago, their creative director wrote about a new paradigm of Magic design. One of their solutions was to pay careful attention to how they spent complexity points in ways that affected new players.

Three examples:

1. Common Cards

In Magic, when you buy a new pack, 11 cards are "common", 3 are "uncommon" and one is "rare". Experienced players buy lots of cards and can have access to lots of rares, but new players generally just buy a few cards, so most of their cards are common. Therefore, the complexity of the cards at common determines how much complexity newcomers have to deal with.

2. Keywords

One way to reduce "effective complexity" is to bundle concepts together in a keyword. Instead of saying "this creature deals damage to each of the creatures blocking it and then deals the remainder of its damage to the player", it just says "Trample". There's an initial cost of learning what Trample means, but afterwards, every time you see the word "Trample" on a creature it works the same way.

Trample has some neat things going for it: it sounds evocative, and gets to build off of existing ideas in your brain. You already know what a big animal looks like. You can imagine a small creature getting in the way of the elephant, and it slowing the elephant down slightly but not really stopping it, and the elephant continuing on, trampling over it, and then going on to attack some bigger target.

This imagery is helpful for intuiting what the rules mean, even if the wording is somewhat confusing.

The problem comes when you introduce too many keywords at once. It gets overwhelming. Which brings to a final concept:

3. Evergreen keywords

Every 3 months, new magic cards are released to keep things fresh. New keywords are introduced (usually 3-5).

But there are some keywords (like Trample) that are *always* in season. There are about 16 evergreen keywords. Many of them are pretty intuitive (such as flying creatures only be able to be blocked by other flying creatures) so they aren't hard to learn.

A new player has an implicit goal of "learn all the evergreen keywords", which is a manageable task.

II. Building a high level conversation

I think some of this applies to the rationalsphere, where a lot of important concepts have been built up, or, combined together from neighboring disciplines. (See Anna Salmon's Single Conversational Locus)

Jargon is *useful*. They let you summarize a complex concept in a single word, and then have deeper conversations where each word packs a lot more meaning.

I have a lot of thoughts about how to do jargon *right*, which are beyond the scope of this post. But to summarize, I think good jargon:

- encapusulates an idea that's important to build off of
- lets you distinguish between *similar* concepts that have importantly-different-nuances. (viral infection vs bacterial infection)
- provides some context clues that help you learn it (the way Trample does), while...
- ...not *also* resulting in people confusing what it means (a bad example perhaps being "negative reinforcement", which is not actually the same thing as "punishment")

Some considerations:

1) Sometimes you want a 101 space where you're either introducing ideas to a broader audience. Sometimes you want a 201 space where you're building on those ideas (either helping somewhat-less-newcomers build up a more advanced understanding, or literally developing new content at the cutting edge)

2) Different venues of conversation can have both different expectations of who-is-participating, and different social norms of what kind of participation is encouraged. (i.e an academic journal, a semi-formal internet forum, a facebook post)

3) Some concepts are pretty standalone: layfolk can learn them and use them immediately without having to fit them into a big edifice of theory

4) Furthermore, some concepts make good "gateway" terminology. They're useful standalone, but then they open up a world of ideas to you that you can then further explore.

So my thought is basically: if you are developing jargon, pay extra attention to whether this is Common or Expert Level jargon. There's not a clear dividing line between them, but roughly:

Common Jargon means you're expecting it to be a useful enough idea for layfolk to use regularly (or, you'd like to be able to have conversations with layfolk, or write popularization articles, that rely on the term already percolating into the mainstream, or, use it as a gateway term)

Consequently, it's much more important to put a lot of effort into choosing a term that:

- resonates easily, is memorable...
- ...but avoids people latching onto the wrong aspect of it and misinterpreting it
- doesn't sound like a weird insider term...
- ... but maybe ideally hints at a broader ecosystem of ideas

Expert Jargon is only really useful if you're buying into a broader ecosystem of ideas that build on each other. Accessibility and avoiding misunderstanding is still important if possible but being precise and build-on-able is more valuable.

Further Reading

This post was inspired by and builds upon:

Complexity is bad (Zvi Mowshowitz)
The Purple Sparkly Ball Thing (Malcolm Ocean)
New World Order (Mark Rosewater)

Notes From an Apocalypse

(This is a loose adaptation of a talk I sometimes give on the Cambrian Explosion, smoothed a bit for popular consumption. That talk, in turn, draws heavily from a 2006 paper by Charles Marshall, titled “Explaining the Cambrian ‘Explosion’ of Animals”. It can be read in full here. I’m mostly just trying to test out the new site with an essay I had lying around, be moderately entertaining, and maybe try to suggest this event as a topic of interest for those who care about intelligence explosions and cognitive emergence. If I succeed in all three, then I encourage you to start with Marshall for the more technical, thorough, and correct analysis.)

A.

The Cambrian Explosion is our name for an event that took place about 540 million years ago, one that accounts for the sudden appearance of advanced animal life. Emphasis on “sudden”; it’s like somebody flipped a light switch. What follows is just a story (fragments of a story, really), but it’s a pretty good one, and I don’t think it’ll lead you too far astray.

Let’s start 540 million years ago. The planet is still recovering from a series of catastrophic “snowball Earth” phases, ice ages so severe that the oceans had frozen right across the equator; the tropical glaciers wound their way through loose sediment to leave an enduring footprint. Now that the last of these has thawed, it’s finally starting to warm up in a more permanent way and get a little more hospitable. There are two major continents at the moment, but one of them is starting to break up into smaller and smaller chunks. As with our rising temperatures, this helps create a more habitable planet, since it gets you more high-nutrient coastlines per unit land area. Importantly, the last hundred million years have provided us with oxygen concentrations in the atmosphere at basically 21st-century levels. To the unprepared organisms that dominate the first half of Earth’s history, oxygen can act as a hazardous toxin, but if you harness it correctly you can really kick your metabolism into a higher gear. A good time to be alive, really.

The land is mostly barren. A little bit of pond scum holding on in shallow pools and other moist areas, maybe, but seeds and placentas won’t exist for a long age yet, and so far every reproductive system requires the presence of standing water. The ocean, though, the ocean is a different story, full of life and living. If you go by the rock record, the most common form of life is the stromatolite, a colony of green plant-microbes that secrete stone in distinctive whorling patterns. (As always, history most remembers those who write history). They flourish in shallow and sunlight-touched waters, gradually constructing ranges of gently rolling hills that are each sometimes meters across. Larger (but still microscopic) Eukaryotes wander through these hills like grazing deer. Or maybe like wolves. Unlike the helpful stromatolites themselves, most of these creatures rarely leave informative fossils, so it’s hard to say how complicated the ecosystem actually is. Certainly it’s a rich and dynamic biological scene, but to my mind it always feels just a little bit like Hobbiton. Gentle with a hidden strength, simple enough and flexible enough to endure. The official name for this time period is the Proterozoic, but when geologists think nobody is listening, we sometimes call it the Boring Billion.

But here in the closing years of that era, as temperatures thaw and the coastlines unwind and the atmosphere fills with oxygen, it’s starting to get a bit more

interesting. Big. Some of the Eukaryotes are behaving oddly, bunching up into colonies that are a bit better at finding nutrients or sunlight. Often, they give up roaming and gather together in big fronds, anchoring themselves to the seafloor and spreading the sail out broadside to catch as much nutrient-rich ocean water as possible as it flows by. Sporifera, the sponges, use a different strategy, pulsing flagella inwards to create little artificial currents that channel nutrients towards themselves. Cnidaria, the jellyfish, actually lift off from the ocean floor altogether and drift through the water looking for scraps among the drifting plankton. The cells in these colonies are starting to take on specialized roles, this one binding the group to the soil, that one a tentacle, but their overall simplicity falls well short of what we usually mean by words like 'animal.'

Despite these omens, it's all rather sedate. Not much seems to change day to day, millenium to millenium. You'd be forgiven for checking your watch a few times every epoch. You might even take a nap, stay in bed for a few million years and relax. Maybe twenty million if you hit the snooze button. But when you wake up, you'll be in for a shock.

B.

Somebody *broke it*. They scourged the Shire. In just a few million years, no more than a wink of geological time, everything has changed. Many of those ever-green stromatolite hills have been stripped bare, and the ones that remain are going fast. Giant monsters prowl through the wreckage, hunting for anything made out of complex organics, hunting each other. There is a bewildering array of new lifeforms. Some of them are covered in spikes, others with odd numbers of eyes and strange probing tentacle-mouths. All of them have elaborate organ systems, strange tissue masses that express themselves in radically different, interlocking ways, despite having the same genome. The rise in diversity, and in disparity, is unequaled by any other moment in Earth's history. Something like *half* of all 21st century animal phyla trace their origins back to that brief moment of generation. Take all the creative power of the last five hundred million years of animal evolution, compress it down to a fraction of a geological instant- that's the power of the Cambrian Explosion. In less than twenty million years, there are molluscs squirming like modern sea urchins, echinoderms clinging to rocks like modern starfish. Even the trilobite, that ancient symbol of ancient life, suddenly appears here fully-formed. And then, swimming through the open waters, you'd see the most surprising thing of all: one of them has a brain. Animalia Bilateria Chordata, the chordate.

This is it, you see. The moment of encephalization, the moment that the cosmos wakes up.

There was the Earth, a giant rock drifting quietly through space. And one day it just spontaneously grows a brain, for no damn reason at all. What kind of rock does that, exactly? What kind of universe?

C.

No, really, why did this happen? What was the mechanism, the bridge that takes us from the boring billion to the era of minds and monsters?

I don't know.

Really, I don't. There are hypotheses, sure. A few scraps of almost-understanding. And plenty of guesses, some of them really good. But the stone can only tell us so much,

and it was such a very long time ago. The all-important middle of our story, the one that gives us the first moments of emerging biological consciousness, has not yet been recovered.

Can it be recovered, even in principle? Harder question.

Darwin, poor Charles Darwin, floundered on these shoals with the rest of us. For him, the only explanation was that we must be missing some huge fraction of the rock, so that it only seems like they all show up at once:

> I cannot doubt that all the Silurian trilobites have descended from some one crustacean, which must have lived long before the Silurian age....Consequently, if my theory be true, it is indisputable that before the lowest Silurian strata was deposited, long periods elapsed, as long as, or probably longer than, the whole interval from the Silurian to the present day.....The case must at present remain inexplicable; and may be truely urged as a valid argument against the views here entertained.

He was wrong. Isotopic dating methods have since confirmed that there is no gap in the rock between the Proterozoic and the Cambrian, no space of hundreds of millions of years for us to move gradually from one extreme to the other. (In this case, we're also early enough in the history of geology that precise chronology was still ambiguous- that's why Charles refers here to the Silurian, a later era.)

And so, as the man says, the Cambrian Explosion may be truly urged as a valid argument against the views which Mr. Darwin entertained. What's at stake here is not just the question of how this process of encephalization first occurred, but also why our most foundational biological theories fail so spectacularly to anticipate it. Are we even asking the right questions?

D.

There's a slightly more modern version of Darwin's first tentative attempts at a patch, one that might explain why so many fossils would be missing from the pre-Cambrian record. If it's true, there never was a Cambrian Explosion, the paragraphs above don't correspond to the world as it was, and the Earth's encephalization took place very gradually, over eons. A relief in some ways (our theory of evolution remains sound), but a tragedy in others (since we'd probably never be able to peer at what might be the most important moment in the history of life).

Here's a different story:

By the end of the Proterozoic, there has been a thriving multicellular ecosystem for hundreds of millions of years, full of complex animals in a thriving dance of grazing and predation, reproduction and survival, gradually expanding and exploring the space of possible forms. But these organisms are all soft-bodied, with no bones or shells or rigid frameworks of any kind. When they die, their bodies rot immediately, leaving no trace for paleontologists to find. But remember when I mentioned that one of the continents is breaking up? All this new coastline, and new weathering, bring new minerals into the oceans. Suddenly, the system is flooded with dissolved calcium salts, ready and able to be incorporated into bones and other biological machinery. Naturally, a number of different well-established species take advantage of this. Skeletons, spines, and shells become very popular. And when they die, they're preserved, some for a long enough for a poor foolish scientist to stumble across. And from our perspective, there's a bright line with animals on only one side of it.

This is a possibility that we should take very seriously. A good chunk of the scientific community certainly does.

One of the things that swayed these scientists is a method of using ‘molecular clocks’ to learn from living genetic sequences as if they were a sort of fossil. The trick is this: look for genetic sequences of a very specific sort, those that change randomly, protected from the directionality of evolutionary selection pressures, and which have been doing so at a slow and steady rate for hundreds of millions of years. (In practice, certain structural elements of hemoglobin work well.) Sequence this area in two very different animal species for which we have already discovered the age of their last common ancestor- and by measuring the difference, we can precisely calibrate the rate of change.

Then, all you have to do is apply this same procedure to, say, a mollusc and a chordate. If the Cambrian Explosion happened like we say it happened, you should get about 500 million years of genetic drift.

Actual answer? 800 million. This gives us a full 300 million years to go from primitive sponges to trilobites, an eminently reasonable wait.

Still, I feel a little squidgy about this line of reasoning. Molecular clocks are a dangerously fragile tool, for one. And remember also that the 800 million year figure would have complex animals surviving the snowball Earth periods, where a frozen surface layer prevented gas exchange between the oceans and atmosphere; those oceans were anoxic, lacking any oxygen for the animals to breathe. But my main objection is actually a bit simpler: no ichnofossils.

An ichnofossil, or ‘trace fossil’ is just any biological remnant that doesn’t involve the actual biological thing itself. Footprints, burrows, etcetera. They’re kind of a pain to find- trace fossil hunting famously takes place at dawn so that the shadows throw your field area into sharper relief- but preserve a lot of valuable information that actual bones do not. Consider the famous archaeopteryx fossil that preserved, not just the skeleton of a dinosaur like so many others, but also the clear imprint of its feathers. That is the weight of an ichnofossil.

As you might imagine, we’ve really scoured the strata around the Cambrian transition as best we can. For the simpler organisms, we do find a number of trace fossils. That’s why I can tell you about the sponges and jellyfish; we have the impressions they left in the ground as they died. We even have enough detail to learn astonishing things like: “Modern jellyfish are quadrilaterally symmetrical, a circle with for identical quadrants. But when they first emerged, at least some jellyfish were trilaterally symmetrical, a circle with three identical parts, instead.”

But aside from the jellyfish and those weird asymmetrical fronds, we genuinely don’t see much. No burrows, no tracks, no trilobyte-shaped impressions in soft mud. So if there was this huge dynamic ecosystem, why didn’t it leave any footprints, even though conditions were favorable enough for such things that we can find the final resting place of a 600 million year old jellyfish? The most provocative traces we find are narrow (millimeter) horizontal trails or burrows, just 2-3 million years before the Cambrian. That’s the only prelude we’ve found so far, and it’s a brief one.

I don’t think this is strong enough to break the missing-calcium theory outright; if nothing else, it’s an argument from absence. But I do think it’s a pretty important nail in the coffin, and a reason for skepticism. And so- we still have at least a pretty good

reason to think that the Cambrian Explosion is a real thing, and that our search for the middle part of that story is not foolish.

I can think of a few places to look.

E.

Perhaps the answer we're looking for is, 'oxygen'. As we observed earlier, this is a period in Earth's history in which oxygen had increased to near-modern levels somewhat recently.

Oxygen isn't quite an absolute prerequisite for complex food webs, but it's pretty close. Animal metabolism derives energy from a flow of electrons as they move from one molecule to another, almost but not entirely unlike a water wheel powered by a flowing river. The faster that water flows, the more energetic a machine you can power. To do that, you need a source of electrons, preferably in some dense high-energy package like sugar. But you also need to provide a grounding, a place for them to go, something that pulls in electrons as hard as possible. That is, our power system also needs an electron sink. Oxygen does this job remarkably well- thus, as animals, we survive by eating and breathing.

Without oxygen, you need to rely on a less energetic sink. There are species of iron that work alright, and there are some weird hacks available with hydrogen, but the machines that you can make with these power sources are only so impressive. If all you have is iron as an electron sink, you can probably manage grazing pretty well, and you can buy some wiggle room as long as you're microscopic. But something high-energy like multicellular predation is a big ask. At least today, whenever we see low-oxygen environments in the deep ocean and so on, we always see a corresponding reduction in food web complexity and species diversity.

So could a quantitative increase in oxygen levels have produced a qualitative change in the structure of animal life? Possibly. But there was a significant lag between the rise of oxygen and the rise of animal complexity, almost a hundred million years. It begs the question, why not sooner? Oxygen probably had *something* to do with this, but it seems more like a prerequisite than a cause- and we have yet to account for the morphological changes. Why not just use the energy to make bigger fronds? Fronds for miles, fronds to span mountains, fronds beyond the wildest dreams of frondkind!

F.

Think back to those horizontal burrows that show up in the very last days of the Proterozoic. There are a couple ways that this is really, really exciting. First, it implies that sensation probably was starting to get concentrated on one end of the animal, or at least one side- taste, smell, even vision clustered around a single area. That feels suspiciously like we're starting to get nerve clusters that you can almost call a brain.

But also, significantly for our purposes, this means that animals might be starting to develop sophisticated hox genes.

The hox genes are, more or less, the standard library for the structural assembly of animal bodies. It's a DNA-modifying type of gene, one that activates specific other regions of DNA during early development. Need a leg? Activate the 'leg' hox gene, and that will start a huge cascade of related processes that make a torso segment that contains a couple legs plus all the necessary hip-joints and such in a somewhat standardized way.** This makes it easier than you'd think to adapt animal body plans

on the fly; rather than reinventing legs from the ground up every time you want to adapt from quadruped to hexaped, you can just have a mutation that calls the hox gene two more times. It is also a primary mechanism of directionality, in which animal bodies have a clear orientation with a front and back.

As you might imagine, animals almost never survive a mutation to the hox genes proper, let alone thrive and speciate. It tends to mean that you get born without a head or something. So, both humans and house flies tend to have a very similar set. That holds true across the entire animal kingdom, with only a few exceptions- hox genes are frozen in time, proportionate with their importance. Care to guess which types of animal lack fully formed hox genes?

That's right: sponges and jellyfish. Even those have a rough proto-hox thing going on, but it's a far cry short of ours.

It goes without saying that this will have implications for species diversity in the animal kingdom. Adaptation isn't as easy as building an animal out of legos, but it has at least gotten a lot easier. Animals can experiment with body shapes in more radical ways, and be successful more often when they do so- they're now exploring a much larger space of possibilities.

We're starting to piece together a rough framework here, in which the rise of oxygen and the slow development of meta-genomic advantages work together to provide space for a more dynamic ecosystem, but is that enough to make sense of something as dramatic and surprising as the Cambrain Explosion? It's a start! But let's see if I can't make it a little more complicated.

**I am lying harder than usual right now. This is complicated and I am not a geneticist. Please never believe me about anything.

G.

We've got the oxygen levels needed for large-scale predation and multilayered food webs, and we've got the genetic toolkit to move quickly through different animal shapes as evolutionary pressures come down on us, sure. But again, why so *many*, and why all at once? What splintered the animal kingdom so thoroughly, and spread the shards of it so widely? Before the Explosion, we apparently had two or three general body plans, each with an accompanying niche. Afterwards- everything else.

Personally, my favorite answer to that question is, 'eyeballs'.

Maybe also noses and ears, I guess. But those are a little less directional, and we're pretty sure that eyes did in fact develop around this time.

Either way, what I really mean is, 'long distance detection of nutrients'. This synchronizes nicely with the directional motion that shows up around this time, the 'front and back' and 'top and bottom' innovations that allow an animal to see food and then move towards it.

What I *really* really mean is, 'predation.'

On the one hand, this makes it a lot more viable to be one step up in the food web. A blind predator is going to have some trouble; the ability to see and pursue opens up a new niche. But only the *one* new niche, so that alone still doesn't explain the riotous diversity we new encounter.

Consider the criteria you must satisfy to be a successful precambrian animal. You're going to need to absorb as many complex carbon molecules as possible. You're going to have to solve the reproduction problem somehow. And you're going to have to be structurally sound, rather than collapsing under your own weight or something. It's fairly simple, mostly revolving around being able to access as much seawater as possible so you can filter organics out of it.

And in fact, the three major solutions we see in the fossil record are, "have a large broadside and catch water as it moves by", "actively pump water towards yourself", and "move quickly through the water." When you think about it, that's a fairly exhaustive list. Every one of these forms is clearly exploring ways to have physical contact with as much seawater as possible, and almost nothing else. And there are only so many ways to be the best at that job.

But now let's add another criterion: "Don't get eaten."

It's not just that this is a fairly flexible and ambiguous constraint. It's that we've lost our chance to monomaniacally focus on the one strategy of 'contact lots of water'. And when you have to balance different needs that are in competition, there are a few different trends that tend to emerge from that process. First, none of your strategies will be quite as successful as they were when you didn't have to make any tradeoffs. And second, there will tend to be a number of different ways to balance those needs, each of which is roughly as effective as the others.

This is a really important but kind of abstract point, so I'm gonna hammer at it for a little bit. *As the number of constraints increases, the number of different 'best' solutions will tend to increase combinatorically.* If you're a proto-sponge, and you're suddenly under all this pressure from being hunted, you might try covering yourself in spikes, or borrowing into the mud, or growing in nasty inhospitable places where predators don't want to go, or just picking up a certain amount of mobility. But remember, you still have to worry about filter-feeding yourself. And the proto-jellyfish has a similar number of different options, and the weird frond-looking things too. The net result is that we end up with a really huge number of equally viable body plans, rather than the two or three that are the winners of a simpler problem.

Basically, the idea here is that predation was such a radical change to the environment that it fractured a small number of deep ecological niches in to a large number of somewhat shallower niches. The radical speciation of the Cambrian Explosion is simply the natural result of those niches being explored and filled.

H.

And somehow, from that crucible, the brain.

We aren't there yet. We know that long-distance senses like vision probably would have given rise to directional motion and directional anatomy, and that the combined needs of directional motion and of sensation managed to concentrate nerves in one particular region of the animal. Maybe we can find out more with detailed genetic analyses, or maybe we'll hit the motherload with some hugely important fossil discovery. Hard to say, but I doubt we're done yet.

Analogies are always perilous, and I will personally get very annoyed at anyone who tries to make a political metaphor out of this or something. But in more general terms, it's worth taking a step back and thinking about what the Cambrain Explosion says

about our universe. The run up to this apocalypse seem to have included at least a few generalizable events:

The first is a deep well of underexploited potential energy. Oxygen, in our case, meant that the ratio between the possible and the actual was somewhat larger than usual. Finding a metabolic pathway to exploit oxygen was difficult, but once the blueprint existed, the machinery itself wasn't particularly difficult.

The second is that new layers of useful abstraction emerged, which allowed innovation and conceptual mobility on larger scales than had previously existed. We accelerated faster through our search space.

These do not themselves lead directly to an abundance of new forms. Rather, the fits and starts of the early successes present new challenges to existing institutions. Those organizations that successfully adapt to the new environment are still constrained by the additional complexity, and windows open for entirely novel forms.

And that's about as much of a metaphor as I'm willing to make. Still, it probably influences a lot of the way I think about, e.g., Silicon Valley. And it's probably important to remember that any Singularity would be the *second* intelligence explosion that the Earth has gone through.

The stromatolites are still around, by the way. They never conquered the world again, but they still build their little hills in hypersaline and hyperthermal waters, places where animal life can't survive. There's a nice cluster of them in the Bahamas, which makes for some nice field expeditions for geobiologists. It's not a bad retirement.

Against EA PR

Scott Alexander recently wrote about weird effective altruism. Many people (mostly, but not entirely, people who aren't effective altruists) offered the opinion that weird effective altruists should be banned from EA, or at least shouldn't be allowed to give talks at EA Global and have blog posts written about them. Weird effective altruist causes are (sort of by definition) off-putting to most people; therefore, if you want people to donate to global poverty relief, you should kick out all of those people concerned about farmed animal welfare/AI risk/wild animal welfare/psychedelics research/suffering in fundamental physics, lest we scare the normies.

There are many reasonable critiques of this point of view, including that it's not remotely clear that any of those claims are more frightening to normal people than "it is morally obligatory to make personal sacrifices in order to help poor, faraway black people." But ultimately I reject the entire premise.

I'd like to be clear about what I'm not saying in this post. I am not saying all "weird effective altruism" causes are effective; I believe some are and some aren't. I think many effective altruists are not taking seriously enough the difficulty of figuring out how effective highly speculative causes are, and that unless we seriously address this we're going to waste potentially millions of dollars on boondoggles. And I suspect a lot of weird effective altruism tends to over-explore certain cause areas (for example, things you think of if you read too many science fiction novels) and underexplore other cause areas (for example, boring things). I don't intend this post to be a whole-hearted defense of weird effective altruism, but simply a criticism of a single narrow argument too often wielded against it.

So the question arises: why is effective altruism a thing at all?

Most people care about charity effectiveness, at least a little bit. They look up their charities on Charity Navigator before donating; they object to money being spent on big CEO salaries or on overhead instead of on services; they circulate criticisms of the Susan G Komen Foundation and PETA. And yet not only do most social programs not work, for the vast majority of programs we simply haven't collected the information to see whether it works or not. This isn't a "no one cares about starving Africans" thing; the state of the evidence on warm-fuzzies American medical and educational interventions is equally poor.

Part of the problem is that while people care about effectiveness some, they don't care about effectiveness that much. They are willing to google a charity to see whether it is an outright scam, but they're not willing to read academic papers to see if the charity's intervention works. They're definitely not going to put in the time to separate intuitive but misleading measures of effectiveness (CEO pay) from actually good measures of effectiveness (randomized controlled trials).

The other part of the problem is that all charity advertisements are a hellhole of epistemic doom and despair.

Let's pick on Feeding America. Not because it's an unusually bad charity (it's not), but because it's large and typical.

Looking on their webpage, I find out immediately that 1 in 8 Americans struggles with hunger. That sounds awful! After clicking through several pages, I find that the source

is this document, in which 1 in 8 households (not individuals) are food insecure. You can click through to the document to read the full operationalization of food insecure (it's on pages 3-4). Food insecure households include, for instance, a household that sometimes worries about whether they'll run out of food, feeds their children only a few kinds of low-cost food to avoid running out of food, and sometimes can't afford to eat balanced meals. While obviously this household is experiencing a good deal of suffering and Feeding America can help them, it's not exactly what the average person would think of when they hear the word "hunger." This is actively misleading.

I click through to Our Work, where I learn that Feeding America has fed four billion meals last year. What percentage of people who would otherwise go hungry did they feed? 10%? 50%? 99%? How many of their meals went to people who would have otherwise gone hungry, versus people who would have been able to figure out some other way to get enough to eat? Feeding America does not provide any insight into these important questions.

98% of all donations raised go directly to helping people in need: according to Charity Navigator, this refers to program expenses, with 1.1% of their income being spent on fundraising and 0.3% spent on administrative expenses. Would increasing their percent spent on fundraising allow them to help more people by raising more money? Would increased administrative expenses, say, reduce the amount of food waste by hiring someone to improve their distribution practices? We simply don't have enough information to know.

In short, Feeding America is misleading about the scope of the problem they're dealing with and does not provide the necessary information to assess their effectiveness in dealing with it.

Again, I am not picking on Feeding America because it is bad. The reason that charity is a total epistemic hellhole is that all charities are like this. The beloved effective altruist charity the Against Malaria Foundation explains on its homepage that 100% of donations go to buy nets (because presumably in a perfect world AMF employees would not need to earn a salary to pay for such luxuries as "homes" and "food") and entirely omits the fact that most nets will not actually prevent any cases of malaria.

Of course, I'm being unfair here. The purpose of a charity's website is not to tell the complete and unvarnished truth, it's to get people to donate. How many people have actually read a GiveWell charity report all the way through without their eyes glazing over by the time they get to "Niger, Burundi, Malawi, and Liberia Prevalence and Intensity Studies"? If the charity actually had a proper cost-effectiveness assessment rather than a bunch of oversimplified bullet points, everyone would get bored and decide to catch up on Game of Thrones instead and no meals or mosquito nets would be bought at all.

And the harm here seems pretty small. So maybe "100% of public donations go to buy nets" means "we got some people to allocate money towards paying our employees instead of towards nets because you're an idiot who thinks nonprofit employees can survive on nothing more than the satisfaction of doing good." So maybe "struggles with hunger" means "at least one member of the family has missed one meal in the past year due to not having enough money and also the children do not eat enough vegetables" instead of "is hungry most of the time." It's not like they're outright lying, and it's for a good cause. Would you rather people spend that money on a new pair of shoes instead?

But the fact of the matter is that the Red Cross makes the same calculation about disaster relief, and the American Cancer Society makes the same calculation about cancer treatment, and the Smithsonian makes the same calculation about preserving priceless historical artifacts. And that means that it's extraordinarily difficult to figure out really basic questions about charities you might want to donate to, like:

- How much does the problem the charity is trying to solve affect people's lives?
- How many people does the problem the charity is trying to solve affect?
- Does this charity actually help with the problem it is trying to solve?
- If I donate to this charity, will the money go to really important programs that have a big effect on people's lives, or do they already have enough money for all of that and my donation would go to something that doesn't actually do that much good?
- Is this charity better than other charities I might donate to?

Which is the reason effective altruism is possible at all.

As far as I'm aware, effective altruist charity evaluators are the only people who are trying to answer this sort of question for the general public (although presumably some big foundations like the Gates Foundation are trying to answer it for themselves). This is our thing. This is the value we add over a Salvation Army bell-ringer who happens to have some fliers for Idealist.

I don't care about effective altruists' personal honesty. Lie to your parents about your dating life, shade the truth on your resume, compliment your friend's hat which vaguely resembles a dead opossum, whatever. Hell, if you're working for a top charity that isn't an explicitly effective-altruism-branded top charity, do the epistemic hellhole thing. Everyone else is and you might as well try to grab some of the charity budget for things that actually work.

But when you are speaking as an effective altruist-- don't get complicated, don't get clever. Just say what you think the best cause area or charity or career is. Every time you think to yourself "well, I think AI risk is more important, but it'll turn people off, so I should probably say the Against Malaria Foundation," the effective altruism movement takes one more step towards being the same as any other group of charitably-minded nerds.

I go pretty far on this. A lot of introductory effective altruism material uses global poverty examples, even articles which were written by people I know perfectly fucking well only donate to MIRI. I think people should generally either use examples from the cause they actually think is most effective, or use an equal number of existential risk, animal welfare, and global poverty examples, in order to reflect the disagreement in the effective altruist community.

I'm not saying you should pay literally zero attention to public relations. There are lots of things you can do to be more persuasive that don't involve misleading people. You can show people pictures of sad animals or happy African children. You can wear professional clothes offline or write with proper grammar online. You can be kind and respectful and try to see things from other people's points of view. But you must abjure all attempts to persuade people by doing anything other than giving people

your best assessment of all the evidence, including all the nuance and all the caveats, even if it might turn them off.

Strategic Goal Pursuit and Daily Schedules

In the post Humans Are Not Automatically Strategic, Anna Salamon writes:

> there are clearly also heuristics that would be useful to goal-achievement (or that would be part of what it means to “have goals” at all) that we do not automatically carry out. We do not automatically:
>
> (a) Ask ourselves what we’re trying to achieve;
>
> (b) Ask ourselves how we could tell if we achieved it (“what does it look like to be a good comedian?”) and how we can track progress;
>
> (c) Find ourselves strongly, intrinsically curious about information that would help us achieve our goal;
>
> (d) Gather that information (e.g., by asking as how folks commonly achieve our goal, or similar goals, or by tallying which strategies have and haven’t worked for us in the past);
>
> (e) Systematically test many different conjectures for how to achieve the goals, including methods that aren’t habitual for us, while tracking which ones do and don’t work;
>
> (f) Focus most of the energy that *isn’t* going into systematic exploration, on the methods that work best;
>
> (g) Make sure that our "goal" is really our goal, that we coherently want it and are not constrained by fears or by uncertainty as to whether it is worth the effort, and that we have thought through any questions and decisions in advance so they won't continually sap our energies;
>
> (h) Use environmental cues and social contexts to bolster our motivation, so we can keep working effectively in the face of intermittent frustrations, or temptations based in hyperbolic discounting;

When I read this, I was feeling quite unsatisfied about the way I pursued my goals. So the obvious thing to try, it seemed to me, was to ask myself how I could actually do all these things. I started by writing down all the major goals I have I could think of (a). Then I attempted to determine whether each goal was consistent with my other beliefs, whether I was sure it was something I really wanted, and was worth the effort(g).

For example, I saw that my desire to be a novelist was more motivated by the idea of how cool it would feel to be able to have that be part of my self-image, rather than a desire to actually write a novel. Maybe I’ll try to write a novel again one day, but if that becomes a goal sometime in the future it will be because there is something I really want to write about, not because I would like to be a writer.

Once I narrowed my goals down to aspirations that seemed actually worthwhile I attempted to devise useful tracking strategies for each goal (b). Some were pretty

concrete (did I exercise for at least four hours this week) and others less so (how happy do I generally feel on a scale of 1-10 as recorded over time), but even if the latter method is prone to somewhat biased responses, it seems better than nothing.

The next step was outlining what concrete actions I could begin immediately taking to work towards achieving my goals, including researching how to get better at working on those goals (d,e,f). I made sure to refer to these points when thinking about actions I could take, it helped significantly.

As for (c), if you focus on how learning certain information will help you achieve something you really want to achieve and you still are not curious about it, well, that's a bit odd to me, although I can imagine how that might occur. But that is something of a different topic than I want to focus on.

Now we come to (h), which is the real issue of the whole system, at least for me. Or perhaps it would be clearer to say that general motivation and organization was the biggest problem I had when I first tried to implement these heuristics. I planned out my goals, but trying to work on them by sheer force of will did not last for very long. I would inevitably convince myself that I was too tired, I would forget certain goals fairly often (probably conveniently the tasks that seemed the hardest or least immediately pleasant), and ultimately I mostly gave up, making a token effort now and again.

I found that state of affairs unsatisfactory, and I decided what felt like a willpower problem might actually be a situational framing problem. In order to change the way I interacted with the work that would let me achieve my goals, I began fully scheduling out the actions I would take to get better at my goals each day.

In the evening, I look over my list of goals and I plan my day by asking myself, "How can I work on everything on this list tomorrow? Even if it's only for five minutes, how do I plan my day so that I get better at everything I want to get better at?" Thanks to the fact that I have written out concrete actions I can take to get better at my goals, this is actually quite easy.

These schedules improve my ability to consistently work on my goals for a couple reasons, I think. When I have planned that I am going to do some sort of work at a specific time I cannot easily rationalize procrastination. My normal excuses of "I'll just do it in a bit" or "I'm feeling too tired right now" get thrown out. There is an override of "Nope, you're doing it now, it says right here, see?" With a little practice, following the schedule becomes habit, and it's shocking how much willpower you have for actually doing things once you don't need to exert so much just to get yourself to start. I think the psychology it applies is similar to that used by Action Triggers, as described by Dr. Peter Gollwitzer.

The principle of Action Triggers is that you do something in advance to remind yourself of something you want to do later. For example, you lay out your running clothes to prompt yourself to go for that jog later. Or you plan to write your essay immediately after a specific tangible event occurs (e.g. right after dinner). A daily schedule works as constant action triggers, as you are continually asking the question "what am I supposed to do now?" and the schedule answers.

Having a goal list and daily schedule has increased my productivity and organization an astonishing amount, but there have been some significant hiccups. When I first began making daily schedules I used them to basically eschew what I saw as useless leisure time, and planned my day in a very strict fashion. The whole point is not to waste any time, right? The first problem this created may be obvious to those who

better appreciate the importance of rest than I did at the time. I stopped using the schedules after a month and a half because it eventually became too tiring and oppressive. In addition, the strictness of my scheduling left little room for spontaneity and I would allow myself to become stressed when something would come up that I would have to attend to. Planned actions or events also often took longer than scheduled and that would throw the whole rest of the day's plan off, which felt like failure because I was unable to get everything I planned done.

Thinking back to that time several months later, when I was again dissatisfied with how well I was able to work towards my goals and motivate myself, I wished for the motivation and productivity the schedules provided, but to avoid the stress that had come with them. It was only at this point that I started to deconstruct what had gone wrong with my initial attempt and think about how I could fix it.

The first major problem was that I had overworked myself, and I realized I would have to include blocks of unplanned leisure time if daily schedules were going to actually work for me. The next and possibly even more important problem was how stressed the schedules had made me. I had to enforce to myself that it is okay if something comes up that causes my day not to go as planned. Failing to do something as scheduled is not a disaster, or even an actual failure if there is good reason to alter my plans. Another technique that helped was scheduling as much unplanned leisure time as possible at the end of my day. This has the dual benefit of allowing me to reschedule really important tasks into that time if they get bumped by unexpected events and generally gives me something to look forward to at the end of the day. The third problem I noticed was that the constant schedule starts to feel oppressive after a while. To resolve this, about every two weeks I spend one day, in which I have no major obligations, without any schedule. I use the day for self-reflection, examining how I'm progressing on my goals, if there are new actions I can think of to add, or modifications I can make to my system of scheduling or goal tracking. Besides that period of reflection, I spend the day resting and relaxing. I find this exercise helps a lot in refreshing myself and making the schedule feel more like a tool and less like an oppressor.

So, essentially, figuring out how to actually follow the goal-pursuing advice Anna gave in Humans Are Not Automatically Strategic, has been very effective thus far for me in terms of improving the way I pursue my goals. I know where I am trying to go, and I know I am taking concrete steps every day to try and get there. I would highly recommend attempting to use Anna's heuristics of goal achievement and I would also recommend using daily schedules as a motivational/organizational technique, although my advice on schedules is largely based on my anecdotal experiences.

I am curious if anyone else has attempted to use Anna's goal-pursuing heuristics or daily schedules and what your experiences have been.

The Virtue of Numbering ALL your Equations

Epistemic status: This is my strongly hold preference, but I dont know to what extent others agree.

I know of two common styles for numbering equations in a scientific publications.

A) Number ALL equations.

B) Number only those equations that you yourself want to reference to in the surrounding text.

A is strictly better than B

- There is no extra effort to in numbering all equations since LaTeX does it for you.
- It becomes much easier to discuss your paper in text and/or online, and in any other situation where the persons involved can't just point directly at the equation they wants to refeer to.

LaTeX automatically number your equations if you use

```
\begin{equation} … \end{equation}
```

LaTeX does **not** automatically number your equation if you use

```
$$ … $$
```

Here is a bit of code you can use if you think "\begin{equation} … \end{equation}" Is too much to write. Copy paste this

```
\newcommand{\be}{\begin{equation}}
\newcommand{\ee}{\end{equation}}
```

into your tex document, before "\begin{document}". Now you can simply use

```
\be … \ee
```

instead of "\begin{equation} … \end{equation}"

Please number <u>all</u> your equations!

The Outside View isn't magic

The planning fallacy is an almost perfect example of the strength of using the outside view. When asked to predict the time taken for a project that they are involved in, people tend to underestimate the time needed (in fact, they tend to predict as if question was how long things would take *if everything went perfectly*).

Simply telling people about the planning fallacy doesn't seem to make it go away. So the outside view argument is that you need to put your project into the "reference class" of other projects, and expect time overruns as compared to your usual, "inside view" estimates (which focus on the details you know about the project.

So, for the outside view, what is the best way of estimating the time of a project? Well, to find the right reference class for it: the right category of projects to compare it with. You can compare the project with others that have similar features - number of people, budget, objective desired, incentive structure, inside view estimate of time taken etc... - and then derive a time estimate for the project that way.

That's the outside view. But to me, it looks a lot like... induction. In fact, it looks a lot like the elements of a linear (or non-linear) regression. We can put those features (at least the quantifiable ones) into a linear regression with a lot of data about projects, shake it all about, and come up with regression coefficients.

At that point, we are left with a decent project timeline prediction model, and another example of human bias. The fact that humans often perform badly in prediction tasks is not exactly new - see for instance my short review on the academic research on expertise.

So what exactly is the outside view doing in all this?

The role of the outside view: model incomplete and bias human

The main use of the outside view, for humans, seems to be to point out either an incompleteness in the model or a human bias. The planning fallacy has both of these: if you did a linear regression comparing your project with all projects with similar features, you'd notice your inside estimate was more optimistic than the regression - your inside model is incomplete. And if you also compared each person's initial estimate with the ultimate duration of their project, you'd notice a systematically optimistic bias - you'd notice the planning fallacy.

The first type of errors tend to go away with time, if the situation is encountered regularly, as people refine models, add variables, and test them on the data. But the second type remains, as human biases are rarely cleared by mere data.

Reference class tennis

If use of the outside view is disputed, it often develops into a case of reference class tennis - where people with opposing sides insist or deny that a certain example

belongs in the reference class (similarly to how, in politics, anything positive is claimed for your side and anything negative assigned to the other side).

But once the phenomena you're addressing has an explanatory model, there are no issues of reference class tennis any more. Consider for instance Goodhart's law: "When a measure becomes a target, it ceases to be a good measure". A law that should be remembered by any minister of education wanting to reward schools according to improvements to their test scores.

This is a typical use of the outside view: if you'd just thought about the system in terms of inside facts - tests are correlated with child performance; schools can improve child performance; we can mandate that test results go up - then you'd have missed several crucial facts.

But notice that nothing mysterious is going on. We understand exactly what's happening here: schools have ways of upping test scores without upping child performance, and so they decided to do that, weakening the correlation between score and performance. Similar things happen in the failures of command economies; but again, once our model is broad enough to encompass enough factors, we get decent explanations, and there's no need for further outside views.

In fact, we know enough that we can show when Goodhart's law fails: when no-one with incentives to game the measure has control of the measure. This is one of the reasons central bank interest rate setting has been so successful. If you order a thousand factories to produce shoes, and reward the managers of each factory for the number of shoes produced, you're heading to disaster. But consider GDP. Say the central bank wants to increase GDP by a certain amount, by fiddling with interest rates. Now, as a shoe factory manager, I might have preferences about the direction of interest rates, and my sales *are* a contributor to GDP. But they are a tiny contributor. It is not in my interest to manipulate my sales figures, in the vague hope that, aggregated across the economy, this will falsify GDP and change the central bank's policy. The reward is too diluted, and would require coordination with many other agents (and coordination is hard).

Thus if you're engaging in reference class tennis, remember the objective is to find a model with enough variables, and enough data, so that there is no more room for the outside view - a fully understood Goodhart's law rather than just a law.

In the absence of a successful model

Sometimes you can have a strong trend without a compelling model. Take Moore's law, for instance. It is extremely strong, going back decades, and surviving multiple changes in chip technology. But it has no clear cause.

A few explanations have been proposed. Maybe it's a consequence of its own success, of chip companies using it to set their goals. Maybe there's some natural exponential rate of improvement in any low-friction feature of a market economy. Exponential-type growth in the short term is no surprise - that just means growth in proportional to investment - so maybe it was an amalgamation of various short term trends.

Do those explanations sound unlikely? Possibly, but there is a *huge trend in computer chips going back decades* that needs to be explained. They are unlikely, but they have

to be weighed against the unlikeliness of the situation. The most plausible explanation is a combination of the above and maybe some factors we haven't thought of yet.

But here's an explanation that is implausible: little time-travelling angels modify the chips so that they follow Moore's law. It's a silly example, but it shows that not all explanations are created equal, even for phenomena that are not fully understood. In fact there are four broad categories of explanations for putative phenomena that don't have a compelling model:

1. Unlikely but somewhat plausible explanations.
2. We don't have an explanation yet, but we think it's likely that there is an explanation.
3. The phenomenon is a coincidence.
4. Any explanation would go against stuff that we do know, and would be less likely than coincidence.

The explanations I've presented for Moore's law fall into category 1. Even if we hadn't thought of those explanations, Moore's law would fall into category 2, because of the depth of evidence for Moore's law and because a "medium length regular technology trend within a broad but specific category" is something that has is intrinsically likely to have an explanation.

Compare with Kurzweil's "law of time and chaos" (a generalisation of his "law of accelerating returns") and Robin Hanson's model where the development of human brains, hunting, agriculture and the industrial revolution are all points on a trend leading to uploads. I discussed these in a previous post, but I can now better articulate the problem with them.

Firstly, they rely on very few data points (the more recent part of Kurzweil's law, the part about recent technological trends, has a lot of data, but the earlier part does not). This raises the probability that they are a mere coincidence (we should also consider selection bias in choosing the data points, which increases the probability of coincidence). Secondly, we have strong reasons to suspect that there won't be any explanation that ties together things like the early evolution of life on Earth, human brain evolution, the agricultural revolution, the industrial revolution, and future technology development. These phenomena have decent local explanations that we already roughly understand (local in time and space to the phenomena described), and these run counter to any explanation that would tie them together.

Human biases and predictions

There is one area where the outside view can still function for multiple phenomena across different eras: when it comes to pointing out human biases. For example, we know that doctors have been authoritative, educated, informed, and useless for most of human history (or possibly much worse than useless). Hence authoritative, educated, and informed statements or people are not to be considered of any value, unless there is some evidence the statement or person is truth tracking. We now have things like expertise research, some primitive betting markets, and track records to try and estimate their experience; these can provide good "outside views".

And the authors of the models of the previous section have some valid points where bias is concerned. Kurzweil's point that (paraphrasing) "things can happen a lot faster than some people think" is valid: we can compare predictions with outcomes. Robin has similar valid points in defense of the possibility of the em scenario.

The reason these explanations are more likely valid is because they have a very probable underlying model/explanation: humans are biased.

Conclusion s

- The outside view is a good reminder for anyone who may be using too narrow a model.
- If the model explains the data well, then there is no need for further outside views.
- If there is a phenomena with data but no convincing model, we need to decide if it's a coincidence or there is an underlying explanation.
- Some phenomena have features that make it likely that there is an explanation, even if we haven't found it yet.
- Some phenomena have features that make it unlikely that there is an explanation, no matter how much we look.
- Outside view arguments that point at human prediction biases, however, can be generally valid, as they only require the explanation that humans are biased in that particular way.

Sabbath hard and go home

Growing up Jewish, I thought that the traditional rules around the Sabbath were silly. Then I forgot to bring a spare battery on a camping trip. Now I think that something like the traditional Jewish Sabbath is an important cultural adaptation to preserve leisure, that would otherwise be destroyed in an urbanized, technological civilization.

Sabbath as easy mode

As a child, I first learned that the Sabbath was a “day of rest,” a day on which you don’t do “work.” I was brought up by liberal Jews in a society in which “work” tends to mean either business or wage labor. Things you do for *money*. Things you do because someone else demands them. This is for the most part how we observed the Sabbath.

But I was also taught about the older traditions in which many categories of mundane activity are forbidden: lighting a fire, cutting or mending cloth, writing or erasing letters. This seemed to me like an arbitrary superstition based on an excessive literality. Surely I could tell for myself whether I was writing as part of a leisure activity or a desk job. Surely I could tell for myself whether I was planting seeds for my private garden, or on a commercial farm. Why avoid these activities in the privacy of one’s own home, doing things for oneself, and not *working* at all?

Likewise, Orthodox Jews must walk to and from their synagogue on the Sabbath, because driving would involve lighting a fire. Automobile engines run on combustion, after all. Liberal Jews often argued, if there is inclement weather, or if the synagogue is far, is it not more *restful* to take an easy drive than to walk?

In short, I thought that the rest of the Sabbath meant, or ought to mean, playing life on easy mode.

Unplugging as leisure

Recently, I’ve been feeling too caught up in local social momentum. When it looked like it would be difficult and take a long time to book a cabin to spend some time alone, I asked a friend to teach me how to go camping, to improve my range of options for solitude, both by directly giving me the affordance for camping, and by more generally expanding the range of living conditions I had experience coping with.

On my first solo two-night camping trip, I forgot to bring a backup battery to charge my laptop or phone. I was car camping, so I could have charged them that way, but I felt like that was outside the spirit of the exercise, and inconvenient anyway. So instead, I mostly kept my phone turned off. Very quickly, I started being able to *think* about aspects of my situation that had been too overwhelming, too *in motion*, to get leverage on the day before. Because I wasn’t *dealing with them*. I wasn’t *keeping up* with anything. I was just present, where I was. I wished I’d done this years ago.

And then I realized: if I had keeping a Sabbath, it wouldn’t have taken *years* to take a step back from social momentum. I’d have gotten a chance within seven days of noticing that there was a problem. And seven days later, another chance, and so on.

Immediately, came the reflexive follow-up thought: of course, not the literal Orthodox Jewish Sabbath. But then I asked my self: why not, exactly?

I went through some of the more onerous-seeming requirements. You are not permitted to write. But when I went on a meditation retreat, they *also* asked us not to write. And I had no problem with that. It did not seem like an arbitrary superstition to me; it seemed like part of the discipline of an integrated mental practice.

Maybe the Sabbath too is a discipline meant to cultivate a particular sort of mental practice.

You are not allowed to light fires on the Sabbath, which means no cooking; you eat what has been prepared in advance. On that same meditation retreat, we were asked not to bring or prepare our own food, but to accept what was served to us. That too felt like a natural part of the practice.

Why had I been so ready to dismiss the Sabbath out of hand? Where did this prejudice come from? It came from my childhood self, who was *assuming alienation of labor*.

Work as keeping up

If you do not assume like a modern consumerist that *work* is what you do *for money*, and *leisure* is what you *spend money* on, then what is work? It is the activity of producing or maintaining the artifacts necessary for the ongoing production of sustenance. It is the activity of keeping up with reality. And in a civilized society with specialization of labor, where your work is only productive because it is integrated with the work of many others, work is the practice of keeping up with the predominant social reality.

What is leisure, then? Leisure is time when you are not responding to a persistent stream of demands. Not your boss, but not a television commercial or newsfeed either. You can take a walk, or sit silently with friends, and let your mind wander.

Leisure is crucial for a very particular sort of freedom. Not freedom as the range of options presented to you, or the absence of overt restrictions on your behavior, but the amount of autonomy you have in practice, the extent to which the choices you are making are determined by the combination of your own preferences and foresight, rather than the result of being led down a path of someone else's design.

The distinction between this sort of work and leisure is not a perfect match to the Sabbath prohibitions.

You can read a book on the Sabbath (which was not allowed at the meditation retreat), and engage with your whole mind, so long as you do not take notes. So long as you do not try to produce some useful artifact, for your future self to pick up and run with.

You can also talk. Jews do not engage in Noble Silence on the Sabbath; it is not a day of silence. But it cuts out some of the more cognitively costly practices of daily life.

Sabbath as hard mode

Some automation plans make sure to include what they call a human in the loop - on some level of abstraction, every decision is reviewed by a human. You can think of the Sabbath as playing life on hard mode in order to make sure that there is a human in your loop.

You would not want to do this sort of thing all the time. But it might make sense to do periodically - perhaps once a week - as a stopgap measure to combat attention drift. If powerful and pervasive cultural forces are out to get you, you ought to check in from time to time with yourself, and other people with whom you have local, high-quality relationships, to give yourself a chance to notice whether you have gotten got for too much.

Daily meditation or reflection practice has something to offer on this front. So does the Quaker practice of silent worship. And so does the Jewish Sabbath.

Sabbath as alarm

One more useful attribute of the Jewish Sabbath is the extent to which its rigid rules generate friction in emergency situations. If your community center is not within walking distance, if there is not enough slack in your schedule to prep things a day in advance, or you are too poor to go a day without work, or too locally isolated to last a day without broadcast entertainment, then *things are not okay*.

In our commercialized society, there will be many opportunities to purchase palliatives, and these palliatives are often worth purchasing. If living close to your place of employment would be ruinously expensive, you drive or take public transit. If you don't have time to feed yourself, you can buy some fast food. If you're not up for talking with a friend in person, or don't have the time, there's Facebook. But this is palliative care for a chronic problem.

In Jewish law, it is permissible to break the Sabbath in an emergency situation, when lives are at stake. If something like the Orthodox Sabbath seems impossibly hard, or if you try to keep it but end up breaking it every week - as my Reform Jewish family did - then you should consider that perhaps, despite the propaganda of the palliatives, *you are in a permanent state of emergency*. This is not okay. You are not doing okay.

So, how are you?

The Iron Law of Evaluation and Other Metallic Rules.

This is a linkpost for https://www.gwern.net/docs/sociology/1987-rossi

Gwern talks about the reasons why most policies that have been evaluated don't actually improve sociological problems. (poverty, dependency, mental illness, crime) I find it surprising they don't hurt either. Gwern hypothesizes that it might be sample bias or maybe the forces underlying these problems are more powerful than the current capabilities of our institutions to handle. (eg. age? genetics?).

It's pretty unsatisfying. I hear that giving aid to the poor is treating symptoms not underlying causes, but we may not actually know the underlying causes to begin with. I want to say maybe spend more money on sociological research instead of services to the poor? Is that just yet another untested "clearly obvious" mass sociological policy?

Stupid Questions - September 2017

The stupid questions thread was one of the regular threads on LessWrong. It's a place where no question is to stupid to be asked and anybody who answers is encouraged to be kind.

This thread is for asking any questions that might seem obvious, tangential, silly or what-have-you. Don't be shy, everyone has holes in their knowledge, though the fewer and the smaller we can make them, the better.

Please be respectful of other people's admitting ignorance and don't mock them for it, as they're doing a noble thing.

Exposition and guidance by analogy

(Below I attempt to reproduce a chart I can't [?] embed, from *Models and Analogies in Science* by Mary Hesse; I don't know how it will display on all screens, but the source is this image at the top of the cross-post version here.)

```
                    At an elementary level we can set up the following
correspondences:
```

```
                    WATER WAVES          |       SOUND        |          LIGHT
Produced by motion  | Produced by motion | *Produced by moving
 of water particles |  of gongs, strings |   flame, etc.
Reflection          | Echoes, etc.       | Reflection in mirrors
Properties of dif-  | Hearing round      | Diffraction through
 fraction           |  corners           |  small slits, etc.
Amplitude           | Loudness           | Brightness
Frequency           | Pitch              | Color
Medium: Water       | Medium: Air        | *Medium: "Ether"
```

In the table, we see a lot of apparent correspondences between water waves, sound, and light. The “horizontal” notion of similarity lets us notice that sound echoes and light reflects, or that these things all have some sort of “intensity” and “flavor”.

But it’s important to introduce the “vertical” analogy—the items in each column are related by some causal, organizing principle, and there’s a correspondence between those principles. We expect all sorts of things to have similar traits entirely by accident. You won’t get very far filling in a table’s gaps by arguing “this is like that”—better to say “this model is like that model”. You’re taking advantage of a ready-made language (with entailed internal relationships) for a metaphoric redescription of a new subject. In this way analogies can be a useful guide to teaching and learning new models in new domains. (You’ll realize, for example, that “produced by moving flame” isn’t really the appropriate correspondence there, because the motion of the flame doesn’t have to do with color in the way that the motion of a gong has to do with pitch, and eventually you’ll learn something about the production of light.) But what’s this about the medium of light—“ether”?

Well, if we observe that the vertical analogy works for the first two models, and it works for the third up to that point, then light having a medium starts seeming plausible enough for us to start looking into it. And it additionally suggests to us how to go about looking. But while the vertical analogy gives us a stronger inductive inference than does the horizontal analogy alone, it’s still quite weak. Light, it turns out, doesn’t seem to propagate through an ether. (But even the “negative analogy”—the apparent hole in the vertical analogy where light’s medium would go—suggests that “why not?” is an interesting question.)

There are two parts to what I just said, so I’ll work them out a little further:

The use of analogy in science is partly pedagogical—it’s about explaining things [as in providing exposition, not reasons or causes] in terms of better-understood things, through their models’ shared structure, or their horizontal points of similarity or

difference (positive and negative analogies). If the structure of the relation you're trying to draw is somewhat confusing on its own terms, or not readily distinguished from a similar model, it could be easier to communicate with reference to another domain. It's easier to understand what we could possibly mean by "light is a wave" if you already know about water waves. And if we're being careful, we say "light is a wave like how water waves are waves," not "light is like water"—we care about communicating the vertical relation, not arguing for horizontal similarity.

And the use of analogy is also about guiding discovery—"neutral" analogies becoming definitely positive or negative as they're used to pinpoint places to investigate. It can be useful to make tentative inferences based on similarity of causal relations to those of a better-understood model, but these inferences really are provisional. You don't know if light has a medium, but the analogy has worked so far, so you design experiments (guided by the analogy) that would detect such a medium. You get a handy working hypothesis and a set of questions to study, and not much else. Your analogy is usually not a very good piece of evidence about its subjects—not good enough to use for engineering—but often still good enough to help decide what's worth investigating. (And, as often when people talk about history and philosophy of science, the big, obvious examples are recapitulated in the everyday work of science on much smaller scales. It's not always about major physical models like electromagnetism and quantum mechanics, but rather implicit in the kind of reasoning that guides investigation from week to week.)

(Philosophers of science also [used to] like to argue about whether analogy is necessary for explanation and/or discovery—that's the dialogue Hesse was participating in above. This is out of scope for us, unless I'm being sneaky.)

Can we take this understanding of analogy outside of science? When is it worthwhile to inject an argument by analogy into your internet? And when is it worthwhile to dispute an analogy?

First, when you need better exposition. If you're making an argument that's hard to spell out in its own language, or easily confused for a more common argument you decidedly don't want to make, then an analogy might be clarifying. (And usually more compelling to read, with that stimulating sensation of insight we get from novel connections, which should make all analogies suspect.) This is where it helps to say "this argument is like that argument" rather than "this is like that". But be careful not to mistake this for substantiating your argument.

Second, to point to questions to investigate. If you're not sure how an argument should come out, you can find other arguments in other domains that look like they flow in the same way. Then the points of analogy are good places to look for the evidence your argument hinges on. And disputing the analogy—saying a point of analogy is neutral or negative—is how your interlocutor points to where they think the contrary evidence lies.

Maybe this is a tedious distinction to keep making, but the usefulness of analogy is not, primarily, in making an inductive inference based on the fact that one model looks mostly like another, where the correctness of that inference depends on the success of the analogy between models. Usually, rather than argument by analogy, you want exposition or guidance by analogy.

Along these lines, it could be generally useful to distinguish between different levels of putting forward and substantiating a claim. You can talk about a position, an argument

for that position, or evidence that the argument hinges on. In doing so you can be anywhere between just pointing to, or describing, or actually demonstrating the bit you're talking about. If someone thinks you're further down the list than you are, then you're liable to get mired in a bad discussion. (Most debates don't get past pointing to where evidence can be found, and most blogging (including this post) doesn't get past pointing to positions or arguments either. Maybe that's fine. Pointing is cheap, both to write and to read. Going deeper can be superfluous, if you're pointing to the obvious. Starting out by pointing could get you to the crucial evidence for resolving a disagreement faster. And so on. [And if you are really just pointing, please consider whether you need so many words.])

In this sense, analogies are for pointing.

Impression track records

It is good to separate impressions from beliefs.

It is good to keep track records.

Is it good to keep separate impression and belief track records?

My default guess would be ‘a bit, but probably too much effort, since we hardly manage to keep any track records.’

But it seems maybe more than a bit good, for these reasons:

1. Having good first impressions, and being good at turning everyone’s impressions into a good overall judgment might be fairly different skills, so that some people are good at one and some are good at the other, and you get a clearer signal if you separate them.

2. We probably by default mostly learn about beliefs and not impressions, because by assumption if I have both and they are different, I suspect the impression is wrong, and so will make me look worse if I advertise that I hold it.

3. Impressions are probably better than beliefs to have track records for, because the point of the track records is to know how much to weight to give different sources when constructing beliefs, and it is more straightforward to know directly which sources are good than to know which aggregations of sources are good (especially if they are mostly bad, because nobody has track records).

As in, perhaps we mostly keep belief track records when we keep track records, but would do better with impression track records. What would we do if we wanted to keep impression track records instead? (Do we already?)

Out to Get You

Epistemic Status: Reference.

Expanded From: Against Facebook, as the post originally intended.

Some things are fundamentally Out to Get You.

They seek resources at your expense. Fees are hidden. Extra options are foisted upon you. Things are made intentionally worse, forcing you to pay to make it less worse. Least bad deals require careful search. Experiences are not as advertised. What you want is buried underneath stuff you don't want. Everything is data to sell you something, rather than an opportunity to help you.

When you deal with Out to Get You, you know it in your gut. Your brain cannot relax. You lookout for tricks and traps. Everything is a scheme.

They want you not to notice. To blind you from the truth. You can feel it when you go to work. When you go to church. When you pay your taxes. It is bad government and bad capitalism. It is many bad relationships, groups and cultures.

When you listen to a political speech, you feel it. Dealing with your wireless or cable company, you feel it. At the car dealership, you feel it. When you deal with that one would-be friend, you feel it. Thinking back on that one ex, you feel it. It's a trap.

Get Gone, Get Got, Get Compact or Get Ready

There are four responses to Out to Get You.

You can Get Gone. Walk away. Breathe a sigh of relief.

You can Get Got. Give the thing everything it wants. Pay up, relax, enjoy the show.

You can Get Compact. Find a rule limiting what 'everything it wants' means in context. Then Get Got, relax and enjoy the show.

You can Get Ready. Do battle. Get what you want.

When to Get Got

Get Got when the deal is Worth It.

This is a difficult lesson for everyone in *at least* one direction.

I am among those with a natural hatred of *Getting Got*. I needed to learn to relax and enjoy the show when the deal is Worth It. Getting Got imposes a large emotional cost for people like me. I have worked to put this aside when it's time to Get Got, while preserving my instincts as a defense. That's hard.

Others make the mistake of *not* hating Getting Got. They might not even notice. This is bad. If you Get Got without realizing, you'll Get Got often for large amounts. Bad

habits will form. Deals won't be Worth It. Reasonable is insufficient: Out to Get You is engineered to fool. Only accept capital letters Worth It.

When you Get Got, *do it on purpose*.

Never Get Got without saying to yourself "I am Getting Got. It is Worth It."

If you realize you've been unwittingly Got, feel sad. Update. Cost is finite, so you should *sometimes* Get Got unaware. It is still unacceptable.

You can choose to Get Got only if you know what you'll be Got for.

You cannot afford to Get Got if the price is not compact.

You can Get Got by a car salesman, saving time and aggravation. Max loss is the price.

You can Get Got with an unlimited phone plan. Max loss is the price.

You can Get Got by a restaurant, club or cruise ship vacation. Leaving money on the table and relaxing could be Worth It, *if you know your max loss and find it acceptable.*

You can Get Got in a relationship. That's the Price of Admission. That's fine *if you know the price and find it Worth It.*

You can buy a AAA game for $60 today rather than $20 next year. Pay $2,000 a year for Magic: The Gathering. Overpay for concert tickets. Wear a symbolic hat. Go vegan. Believe the Knicks will be good next year. If you want. Your call.

There may be no reasonable max loss. Some things want too much.

A clean example is free to play mobile games. If allowed, they charge tens of thousands of dollars. Players called whales are so addicted they pay. The games destroy them.

The motivating example was Facebook. Facebook wants *your entire life*. Users not consciously limiting engagement lose hours a day. Every spare moment is spent scrolling, checking for updates, likes and comments. This reliably makes users miserable. Other social networks share this problem.

An important example is politics. Political causes want every spare minute and dollar. They want to choose your friends, words and thoughts. If given power, they seize the resources of state and nation for their purposes. Then they take those purposes further. One cannot simply give *any* political movement what it wants. That way lies ruin and madness.

Yes, that means *your* cause, too.

This generalizes into most sufficiently intense signaling and status competition. One must always signal harder or seek higher status. This takes over everything you are and eats your entire life. Part of sending sufficiently intense signals is showing that you have allowed this! Maya Millennial has fallen victim. Those keeping up with the Joneses fall victim. Many a child looking fitting in or applying to college falls victim.

Obsession with safety does this.

Television eats people's lives. So do video games. So do drugs and alcohol. One must be careful and know your tenancies and limits.

Ethical arguments do this, ensnaring vulnerable people.

This property is a way to distinguish cults from religions. Cults want it all. Religion wants its cut.

You can only pay off those who charge a bounded price and stay bought. Before you pay the ransom, be sure it will free the hostages.

Would going along result in *cooperation*? Or put blood in the water?

When To Get Compact

Get Compact when you find a rule you can follow that makes it Worth It to Get Got.

The rule must create an acceptable max loss. A well-chosen rule transforms Out to Get You for a lot into Out to Get You for a price you find Worth It. You then Get Got.

This works best using a natural point beyond which lies clear diminishing returns. If no such point exists, be suspicious.

A simple way is a *budget*. Spend at most $25,000 on this car, or $5,000 on this vacation package. This creates an obvious max dollar loss.

Many budgets should be $0. Example: free to play games. Either it's worth playing for free or it isn't. It isn't.

The downside of budgets is often spending exactly your maximum, especially if others figure out what it is. Do your best to avoid this. Known bug.

An alternative is *restriction on type*. Go to a restaurant and avoid alcohol, desert and appetizers. Pay in-game only for full game unlocks and storage space.

Budgets can be set for each purchase. Hybrid approaches are good.

Many cap their charitable giving at 10%. Even those giving more reserve some amount for themselves. Same principle.

For other activities, max loss is about *time*. Again, you can use a (time) budget or limit your actions in a way that restricts (time) spent, or combine both.

Time limits are crude but effective. Limiting yourself to an hour of television or social media per day maxes loss at an hour. This risks making you value the activity more. Often time budgets get exactly spent same as dollar budgets. Try to let unspent time roll over into future periods, to avoid fear or 'losing' unspent time.

When time is the limiting factor, it is better where possible to engineer your environment and options to make the activity compact. You'll get more out of the time you do spend and avoid feeling like you're arbitrarily cutting yourself off.

Decide what's worth watching. Watch that.

For Facebook, classify a handful of people See First. See their posts. No others. Look at social media only on computers. Don’t comment. Or post.

A buffet creates overeating. Filling up one plate (or one early to explore, then one to exploit) ends better.

Unlimited often requires limitation.

Outside demands follow the pattern. To make explanation and justification easier, choose good enough rules that sound natural, simple and reasonable.

Experiments need a chance, but also a known point where you can know to call it quits. Ask whether you can get a definitive negative result in reasonable time. Will I worry I did it wrong? Will others claim or assume I did it wrong or didn’t give it a fair chance?

When to Get Ready

Get Ready when you have no choice.

Getting Ready means battle. An enemy trying to Get You. You are determined not to Get Got. You have done the research. Your eyes are open. You are on alert. You are ready.

You have no choice. The price of surrender is too high. Simple heuristics won’t work. You are already in too deep, or they have something you need and all alternatives are worse.

Sometimes you must accept a bad time and try not to let events get to you. Other times going into battle can be fun. I like games. Games are fun! So are puzzles. Buying a car, planning a vacation, trading for your Magic deck or managing one’s social media interactions can be a game or puzzle. Get the one trying to get you. Get a lot for a little.

There are big downsides.

The game can be fun. The original activity can be fun. *Both at once* is rarely fun. Both means multi-tasking and context-switching, plus a radical shift in emotion and tone. Relaxing into cooperative experience is not compatible with battles of wits and tricks.

The result of this is that you often end up unable to maintain both states at once. Sometimes you end up relaxing, and Get Got. Other times, you focus on not Getting Got and don’t enjoy what you get. Either way, you lose.

The best way out of this is to try and front-load or batch as much of the battle as possible. Sometimes this happens naturally. If you first choose, shop and haggle, then later enjoy the bounty, that’s the ideal way to do battle. Do your best to transform into that sequence, or to make enough choices to transform into a Compact situation.

If this is not possible, consciously switch between modes when needed. Think, “time to pause to not get got,” deal with the issue, switch back. This minimizes bleeding between states. If getting attempts are too continuous, this becomes possible and you need another mode.

You pay for not Getting Got with time and attention. You master arcane details. Time disappears. You spend parties talking tricks instead of living life. If shower thoughts shift to such places, you are paying a high price.

The biggest downside is *you can lose*.

When To Get Gone

Often.

You need good reason to stick around when things are Out to Get You. It is often wise to Get Gone, if you can.

If your instincts say Get Gone, Get Gone. At worst it is only a small mistake.

If your instincts do not say Get Gone, but you can't find a viable approach to another option, Get Gone anyway.

The getting can be insidious. Constant vigilance is required. Many think they can handle it, check all the right boxes and not get drawn in. Some are right. Often they are wrong.

If Getting Got means you lose an order of magnitude bigger than you can win, Get Gone.

If Getting People is how something survives, Get Gone.

Free trial! Automatically renews. Probably won't want? Don't wait. Get Gone.

You **think** you are getting good odds. You are probably wrong.

You **think** you know all the tricks they will try. You are probably wrong.

You **think** there is something is forcing your hand. Make sure this is something you need rather than a want. The word need is thrown around a lot these days.

Getting Gone is worth making sacrifices. Big sacrifices.

If you cannot Get Gone, do not engage more than necessary. Go into Easy Mode. Get what you must. Then Get Gone.

Extensive and Reflexive Personhood Definition

Epistemic status: Speculative idea

It is highly likely that making a friendly AI somehow requires a good definition of what is a morally significant person. The optimal solution may be to figure out what consciousness is, and point to that. But just in case that turns out to be impossible and/or ill defined, we should look at other options too.

In this post I will explore using an extensional definition for this task. Extensional definition is to define a concept by pointing at examples, rather than to describe it in terms of other concepts [1].

Some reasons to be optimistic about using extensional definition for personhood and other important moral concepts:

- This is more or less just (semi) supervised learning, which means we can take advantage of the progress in this field.
- You can teach “I don’t know how to define it, but I know it when I see it” type of information. This means we do not already have know exactly what should be considered a person, at the launch of the AI. We can make it up as we go along and include more and more types of objects in the training data over time.
- The information is not tied to a specific ontology [2].

The main obstacle with extensional definition is that there is no way of making it complete. Therefore, the AI must keep learning forever. Therefore, we need a never ending pool of training data.

* * *

Here is my naive suggestion for person detection system, in the context of how it connects to friendliness [3]:

Hardcoded into the AI:

- The concept of “person” as a process, which as some level of the map, can and should be modeled as an agent with beliefs and preferences [4].
- The belief that persons are better than random at detecting other persons.
- The AI’s goal is to optimise [5] for the aggregation [6] of the preferences [7] of all persons.

Give the AI some sort of initial person detector to get it started, e.g. :

- Program that recognize human faces.
- A specific person that can be queried.

The idea is that the AI can acquire more training data of what is a person by asking known persons for their beliefs. Since the AI tries to optimise for all persons preferences, it is motivated to learn who else is a person.

To begin with the AI is supposed to learn that humans are persons. But over time the category can expand to include more things (e.g. aliens, ems, non human animals [8]). The AI will consider you a person if most previous persons consider you a person. This inclusion mechanism is not perfect. For example, we humans has been embarrassingly slow to conclude that all humans are persons. However, we do seem to get there eventually, and I can't think of any other inclusion method that has enough flexibility.

However, this naive construction for defining personhood is not safe, even if all dependences [3, 5, 6, 7] are solved.

In general, agents already identified as persons, will not agree on who else is a person. This might lead the AI to favour a perverse but more stable to the "who is a agent"-problem. E.g. concluding the persons refers only to the member of a small cult, who strongly believes that they and only they are real people.

It becomes even worse if we consider persons actively trying to hack the system. E.g. creating lots of simple computer programs that all agree with me and then convince the AI that they are all persons.

* * *

[1] Extensions and Intensions

[2] I plan go deeper into this in another post.

[3] Assuming robust AGI.

[4] Note that not everything that can modeled as an agent with beliefs and preferences is a person. But everything that is a person is assumed to have this structure at som level.

[5] I am skipping over the problem of how to line up the state of the world with persons preferences, rather than lining up persons preferences with the state of the world. Part of me believes that this is easily solved by separating preference learning and preference optimisation in the AI. Another part of me believes that this is a really hard problem which we can not even begin to work on until we have a better understanding of preferences [7] are.

[6] I have not yet found a method for aggregating that I like, but I think that this problem can be separated out from this discussion.

[7] The word "preference" is hiding so much complexity and confusion that I don't even know if it is a good concept when applied to humans (see: Call for cognitive science in AI safety). Feel free to interpret "preference" as a placeholder for that thing in a person which is relevant for the AI's decision.

[8] Note that caring about the wellbeing of X is different from considering them persons, in this formulation. If enough persons care about X, then the AI will pick up this preference, even if X is not a person.

Beta - First Impressions

I thought that instead of everyone having to create a separate post for their first impressions, it would be more convenient to create a single post for this discussion. I'll post my own here soon.

The Anthropic Principle: Five Short Examples

(content warning: nuclear war, hypothetical guns, profanity, philosophy, one grammatically-incorrect comma for readability's sake)

This is a very special time of year, when my whole social bubble starts murmuring about nuclear war, and sometimes, some of those murmurers, urging their listeners to worry more, will state: "Anthropic principle."

To teach people about the anthropic principle and how to use it in an epistemically virtuous way, I wrote a short dialogue featuring five examples.

1. Life and death and love and birth

Avery: But how can you *not* believe the universe was designed for life? It's in this cosmic Goldilocks zone, where if you tweaked the speed of light or Planck's Constant or the electron mass by just a few percent, you couldn't have atoms, you couldn't have *any* patterns complex enough to replicate and evolve, the universe would just be this entropic soup!

Brook: I'm not sure I buy the "no complex patterns at all" bit, but even granting it for the sake of argument -- have you heard of the anthropic principle?

Avery: No. Explain?

Brook: We're alive. Life can only exist in a universe where life exists, tautologically. If the cosmos couldn't support life, we wouldn't be having this conversation.

Avery: And so, a universe... created... ngh, I vaguely see what you're getting at. Elaborate?

Brook: And so, a universe created by some kind of deity, and tuned for life, is indistinguishable from a universe that happens to have the right parameters by coincidence. "We exist" can't be evidence for our living in one or the other, because that fact doesn't *correlate* with design-or-lack-of-design -- unless you think that, from *inside a single universe*, you can derive sensible priors for the frequency with which *all* universes, both designed and undesigned, can support life?

Avery: I... oof. That argument feels vaguely like cheating, but... I'll think about it.

2. ...and peace and war on the planet Earth

Avery: In any case, it's interesting that the topic of Earth's hospitality comes up today of all days, given that we so nearly made it inhospitable.

Brook: What do you mean, "today of all days"?

Avery: Oh! You don't know!? September 26 is Petrov Day! It's the anniversary of when some Soviet early-warning system sounded an alarm, and the officer in charge noticed something funny about it and wrote it off -- correctly, as it turned out -- as a false

alarm. If he hadn't, Russia might have "retaliated," starting nuclear war and destroying the world.

Brook: Yikes.

Avery: Yeah! And Wikipedia has a list of similar incidents. We need to be extremely careful to never get into a Cold-War-like situation again: it was incredibly lucky that we survived, and if we get there again, we'll almost certainly melt the planet into a radioactive slag heap.

Brook: Hmm. Nuclear war is definitely bad, and we should try hard to prevent it. But I suspect things aren't as bad as they appear, and that those reports of near-disaster have been exaggerated due to people's credulity for "cool" shiver-inducing things. Theories should get punished for assigning low probabilities to true things: if your model claims that the odds of surviving the Cold War were only 1:1000, it takes a thousandfold probability hit. Any model that predicts Cold-War-survival better, is correspondingly more plausible.

Avery: Not so! Anthropic principle, remember? If the world had ended, we wouldn't be standing here to talk about it. Just as the fact that intelligent life exists shouldn't surprise us (because we can only exist in a universe with intelligent life), the fact that the world didn't end in 1983 shouldn't surprise us (because we can only exist in a world that didn't dissolve into flames).

Brook: I... see what you're saying...

3. Improbability upon improbability

Avery: Oh! And that's not all! According to this article, of the officers who took shifts monitoring that station, Petrov was the only one to have had a civilian education; the others were just taught "to issue and obey orders." It's really lucky that the false alarm went off when it did, instead of twelve hours later when somebody else was at the helm.

Brook: That... rings false to me. I expect there was a miscommunication somewhere between Petrov's lips and your eyes: if there were six officers taking shifts watching the early-warning system, and five of them would've pressed the button, you just declared the probability of surviving this false alarm to be six times smaller: your model takes another 6x hit, just like it would if it *also* claimed that Petrov rolled a die and decided he'd only ignore the warning if it came up 1.

Avery: Anth--

Brook: Don't you dare.

Avery: *coughthropic principlecough*

Brook: *Shut your mouth.*

4. Supercritical

Avery: Fine, fine, sorry. Change of subject: I have a friend who works at the DoE, and they gave me a neat little trinket last week. Here, hold onto this. Careful: it's small, but super heavy.

Brook: Oka-- oh, *jeez*, wow, yeah. What is it?

Avery: A supercritical ball of enriched uranium.

Brook: Gyaah! That's not safe-- wait, *super*critical? That can't be, it would detonate in less than a microsecond.

Avery: And kill us, yes. But we're still alive! Therefore, we must be on the unfathomably tiny branch of possible universes where, so far, when an atom of U-235 in this ball has fissioned, the resulting neutrons tended to *miss* all the other atoms. Thus, the chain reaction hasn't yet occurred, and we survive.

Brook: But Av--

Avery: And you might be tempted to say, "But Avery, that's so improbable I can't even express numbers that small in standard mathematical notation. Clearly this ball is merely platinum or osmium or some other dense metal." But remember! Anthropic principle! You're not allowed to use the fact that you're still alive as evidence! The fact that this ball hasn't detonated is *not* evidence against its being supercritical uranium!

Brook: I-- um. Okay, that is definitely one hundred percent nonsensical sophistry, I just need to put my finger on--

5. Evidence kills

Avery: Sophistry!? I'm insulted. In fact, I'm so insulted that I pulled out a gun and shot you.

Brook: ...what?

Avery: Clearly we're living in the infinitesimally tiny Everett branch where the bullet quantum-tunnelled through your body! Amazing! How improbable-seeming! But, you know, anthropic principle and all.

Brook: NO. I have you now, sucker: even in the branches where the bullet tunneled through me, I would have seen you draw the gun, I'd have heard the shot, I'd see the bullet hole in the wall behind me.

Avery: Well, that all assumes that the photons from my arm and the wall reached your eye, which is a purely probabilistic quantum phenomenon.

Brook: Yes, but still: of the universes where the bullet tunneled through me, in ninety-nine-point-so-many-nines percent of those universes, there is a bullet hole in the wall. Even ignoring the universes where I'm dead, the lack of a bullet hole is overwhelming evidence against your having shot me.

Avery: Is it, though? When you looked at the wall just now, you saw no bullet hole, yes?

Brook: Yes...

Avery: But you, Brook-who-saw-no-bullet-hole, can basically only exist in a universe where there's no bullet hole to be seen. If there *were* a bullet hole, you wouldn't exist -- a Brook-who-*did*-see-a-bullet-hole would stand in your place. Just as the fact that intelligent life exists shouldn't cause you to update (because you can only exist in a

universe with intelligent life), the fact that there's no bullet hole shouldn't cause you to update (because you can only exist in a universe without a bullet hole).

Brook: But seeing a bullet hole doesn't kill me.

Avery: There's nothing *fundamentally different* about death. You wouldn't exist if the world had been destroyed in 1983, but you *also* wouldn't exist if there were a bullet hole: I'd be talking to Brook-who-saw-a-bullet-hole instead. And the fact that you exist can't be used as evidence.

Brook: Are you high on something!? Are you fucking with me!? You're asking me to throw out the whole notion of evidence!

Avery: Oh, yeah, I'm totally messing with you. Absolutely. Sorry. When did I start, though?

Brook: ...sometime between describing Petrov Day -- that part was true, right? Good. -- and telling me that that ball-of-some-dense-metal was made of uranium.

Avery: Correct. But can you be more specific?

Brook: ...tungsten?

Avery: Heh. Yeah, good guess. But I meant about--

Brook: --yeah. I'm not really sure when you started messing with me. And I'm not really sure when you stopped applying the anthropic principle correctly.

Avery: Hmm. That's too bad. Neither am I.

Conclusion

I have no idea whether the anthropic principle is legit or how to use it, or even whether it has any valid uses.

[cross-posted to a blog ; comments here preferred]

Against Individual IQ Worries

[Related to: Attitude vs. Altitude]

I.

I write a lot about the importance of IQ research, and I try to debunk pseudoscientific claims that IQ "isn't real" or "doesn't matter" or "just shows how well you do on a test". IQ is one of the best-studied ideas in psychology, one of our best predictors of job performance, future income, and various other forms of success, etc.

But every so often, I get comments/emails saying something like "Help! I just took an IQ test and learned that my IQ is x! This is much lower than I thought, and so obviously I will be a failure in everything I do in life. Can you direct me to the best cliff to jump off of?"

So I want to clarify: IQ is very useful and powerful for research purposes. It's not nearly as interesting *for you personally*.

How can this be?

Consider something like income inequality: kids from rich families are at an advantage in life; kids from poor families are at a disadvantage.

From a *research* point of view, it's really important to understand this is true. A scientific establishment in denial that having wealthy parents gave you a leg up in life would be an intellectual disgrace. Knowing that wealth runs in families is vital for even a minimal understanding of society, and anybody forced to deny that for political reasons would end up so hopelessly confused that they might as well just give up on having a coherent world-view.

From an personal point of view, coming from a poor family probably isn't great *but shouldn't be infinitely discouraging*. It doesn't suggest that some kid should think to herself "I come from a family that only makes $30,000 per year, guess that means I'm doomed to be a failure forever, might as well not even try". A poor kid is certainly at a disadvantage relative to a rich kid, but probably she knew that already long before any scientist came around to tell her. If she took the scientific study of intergenerational income transmission as something more official and final than her general sense that life was hard – if she obsessively recorded every raise and bonus her parents got on the grounds that it determined her own hope for the future – she would be giving the science more weight than it deserves.

So to the people who write me heartfelt letters complaining about their low IQs, I want to make two important points. First, we're not that good at measuring individual IQs. Second, individual IQs aren't that good at predicting things.

II.

Start with the measurement problems. People who complain about low IQs (not to mention people who boast about high IQs) are often wildly off about the number.

According to the official studies, IQ tests are rarely wrong. The standard error of measurement is somewhere between 3-7 points (1, 2, 3). Call it 5, and that means your tested IQ will only be off by 5+ points 32% of the time. It'll only be off by 10+ points 5% of the time, and really big errors should be near impossible.

In reality, I constantly hear about people getting IQ scores that don't make any sense.

Here's a pretty standard entry in the "help my IQ is so low" genre – Grappling With The Reality Of Having A Below Average IQ:

> When I was 16, as a part of an educational assessment, I took both the WAIS-IV and Woodcock Johnson Cognitive Batteries. My mother was curious as to why I struggled in certain subjects throughout my educational career, particularly in mathematical areas like geometry.
>
> I never got a chance to have a discussion with the psychologist about the results, so I was left to interpret them with me, myself, and the big I known as the Internet – a dangerous activity, I know. This meant two years to date of armchair research, and subsequently, an incessant fear of the implications of my below-average IQ, which stands at a pitiful 94…I still struggle in certain areas of comprehension. I received a score of 1070 on the SAT, (540 Reading & 530 Math), and am barely scraping by in my college algebra class. Honestly, I would be ashamed if any of my coworkers knew I barely could do high school-level algebra.

This person thinks they're reinforcing their point by listing two different tests, but actually a 1070 on the SAT corresponds to about 104, a full ten points higher. Based on other things in their post – their correct use of big words and complicated sentence structure, their mention that they work a successful job in cybersecurity, the fact that they read a philosophy/psychology subreddit for fun – I'm guessing the 104 is closer to the truth.

From the comments on the same Reddit thread:

> Interesting, I hope more people who have an avg. or low IQ post. Personally I had an IQ of 90 or so, but the day of the test I stayed up almost the entire night, slept maybe two hours and as a naive caffeine user I had around 500 mg caffeine. Maybe low IQ people do that.
>
> I did IQTest.dk Raven's test on impulse after seeing a video of Peterson's regarding the importance of IQ, not in a very focused mode, almost ADHD like with rumination and I scored 108, but many claim low scores by around 0.5-1 SD, so that would put me in 115-123. I also am vegan, so creatine might increase my IQ by a few points. I think I am in the 120's, but low IQ people tend to overestimate their IQ, but at least I am certainly 108 non-verbally, which is pretty average and low.

The commenter is right that IQtest.dk usually underestimates scores compared to other tests. But even if we take it at face value, his first score was almost twenty points off. By the official numbers, that should only happen once in every 15,000 people. In reality, someone posts a thread about it on Reddit and another person immediately shows up to say "Yeah, that happened to me".

Nobel-winning physicist Richard Feynman famously scored "only" 124 on an IQ test in school – still bright, but nowhere near what you would expect of a Nobelist. Some people point out that it might have been biased towards measuring verbal rather than

math abilities – then again, Feynman's autobiography (admittedly edited and stitched together by a ghostwriter) sold 500,000 copies and made the New York Times bestseller list. So either his tested IQ was off by at least 30 points (supposed chance of this happening: 1/505 million), or IQ isn't real and all of the studies showing that it is are made up by lizardmen to confuse us. In either case, you should be less concerned if your own school IQ tests seem kind of low.

I don't know why there's such a discrepancy between the official reliability numbers and the ones that anecdotally make sense. My guess is that the official studies give the tests better somehow. They use professional test administrators instead of overworked school counselors. They give them at a specific time of day instead of while the testee is half-asleep. They don't let people take a bunch of caffeine before the test. They actually write the result down in a spreadsheet they have right there instead of trusting the testee to remember it accurately.

In my own field, official studies diagnose psychiatric diseases through beautiful Structured Clinical Interviews performed to exacting guidelines. Then real doctors diagnose them through checklists that say "DO NOT USE FOR DIAGNOSIS" in big letters on the top. If psychometrics is at all similar, the clashing numbers aren't much of a mystery.

But two other points that might also be involved.

First, on a *population level* IQ is very stable with age. Over a study of 87,498 Scottish children, age 11 IQ and adult IQ correlated at 0.66, about as strong and impressive a correlation as you'll ever find in the social sciences. But "correlation of 0.66" is also known as "only predicts 44% of the variance". *On an individual level*, it is totally possible and not even that surprising to have an IQ of 100 at age 11 but 120 at age 30, or vice versa. Any IQ score you got before high school should be considered a *plausible prediction* about your adult IQ and nothing more.

Second, the people who get low IQ scores, are shocked, find their whole world tumbling in on themselves, and desperately try to hold on to their dream of being an intellectual – are not a representative sample of the people who get low IQ scores. The average person who gets a low IQ score says "Yup, guess that would explain why I'm failing all my classes", and then goes back to beating up nerds. When you see someone saying "Help, I got a low IQ score, I've double-checked the standard deviation of all of my subscores and found some slight discrepancy but I'm not sure if that counts as Bayesian evidence that the global value is erroneous", then, well – look, I wouldn't be making fun of these people if I didn't *constantly come across them*. You know who you are.

Just for fun, I analyzed the lowest IQ scores in my collection of SSC/LW surveys. I was only able to find three people who claimed to have an IQ ≤ 100 plus gave SAT results. All three had SAT scores corresponding to IQs in the 120s.

I conclude that at least among the kind of people I encounter and who tend to send me these emails, IQ estimates are pretty terrible.

This is absolutely consistent with population averages of thousands of IQ estimates still being valuable and useful research tools. It just means you shouldn't use it on *yourself*. Statistics is what tells us that almost everybody feels stimulated on amphetamines. Reality is my patient who consistently goes to sleep every time she takes Adderall. Neither the statistics nor the lived experience are wrong – but if you use one when you need the other, you're going to have a bad time.

III.

The second problem is that even if you avoid the problems mentioned above and measure IQ 100% correctly, *it's just not that usefully predictive*.

Isn't that heresy?! Isn't IQ the most predictive thing we have? Doesn't it affect every life outcome as proven again and again in well-replicated experiments?

Yes! I'm not denying any of that. I'm saying that *things that are statistically true aren't always true for any individual*.

Once again, consider the analogy to family transmission of income. Your parents' socioeconomic status correlates with your own at about r = 0.2 to 0.3, depending on how you define "socioeconomic status". By coincidence, this is pretty much the same correlation that Strenze (2006) found for IQ and socioeconomic status. Everyone knows that having rich parents is pretty useful if you want to succeed. But everyone also knows that rich parents aren't the only thing that goes into success. Someone from a poor family who tries really hard and gets a lot of other advantages still has a chance to make it. A sociologist or economist should be very interested in parent-child success correlations; the average person trying to get ahead should just shrug, realize things are going to be a little easier/harder than they would have been otherwise, and get on with their life.

And this isn't just about gaining success by becoming an athlete or musician or some other less-intellectual pursuit. Chess talent is correlated with IQ at 0.24, about the same as income. IQ is some complicated central phenomenon that contributes a little to every cognitive skill, but it doesn't entirely determine any cognitive skill. It's not just that you can have an average IQ and still be a great chess player if you work hard enough – that's true, but it's not *just that*. It's that you can have an average IQ and still have *high levels of innate talent in chess*. It's not quite as likely as if you have a high IQ, but it's very much in the range of possibility. And *then* you add in the effects of working hard enough, and *then* you're getting somewhere.

Here is a table of professions by IQ, a couple of decades out of date but probably not too far off (cf. discussion here):

I don't know how better to demonstrate this idea of "statistically solid, individually shaky". On a population level, we see that the average doctor is 30 IQ points higher than the average janitor, that college professors are overwhelmingly high-IQ, and we think yeah, this is about what we would hope for from a statistic measuring intelligence. But on an individual level, we see that below-average IQ people sometimes become scientists, professors, engineers, and almost anything else you could hope for.

IV.

I'm kind of annoyed I have to write this post. After investing so much work debunking IQ denialists, I feel like this is really – I don't know – diluting the brand.

But I actually think it's not as contradictory as it looks, that there's some common thread between my posts arguing that no, IQ isn't fake, and this one.

If you really understand the idea of a statistical predictor – if you have that gear in your brain at a fundamental level – then social science isn't scary. You can read about IQ, or heredity, or stereotypes, or gender differences, or whatever, and you can say – ah, there's a slight tendency for one thing to correlate with another thing. Then you can go have dinner.

If you don't get that, then the world is terrifying. Someone's said that IQ "correlates with" life outcomes? What the heck is "correlate with"? Did they say that only high-IQ people can be successful? That you're doomed if you don't get the right score on a test?

And then you can either resist that with every breath you have – deny all the data, picket the labs where it's studied, make up silly theories about "emotional intelligence" and "grit" and what have you. Or you can surrender to the darkness, at least have the comfort of knowing that you accept the grim reality as it is.

Imagine an American who somehow gets it into his head that the Communists are about to invade with overwhelming force. He might buy a bunch of guns, turn his house into a bunker, start agitating that Communist sympathizers be imprisoned to prevent them from betraying the country when the time came. Or he might hang a red flag from his house, wear a WELCOME COMMUNIST OVERLORDS tshirt, and start learning Russian. These seem like opposite responses, but they both come from the same fundamental misconception. A lot of the culture war – on both sides – seems like this. I don't know how to solve this except to try, again and again, to install the necessary gear and convince people that correlations are neither meaningless nor always exactly 1.0.

So please: study the science of IQ. Use IQ to explain and predict social phenomena. Work on figuring out how to raise IQ. Assume that raising IQ will have far-ranging and powerful effects on a wide variety of social problems. Just don't expect it to predict a single person's individual achievement with any kind of reliability. Especially not yourself.

Why I am not a Quaker (even though it often seems as though I should be)

In the past year, I have noticed that the Society of Friends (also known as the Quakers) has come to the right answer long before I or most people did, on a surprising number of things, in a surprising range of domains. And yet, I do not feel inclined to become one of them. Giving credit where credit is due is a basic part of good discourse, so I feel that I owe an explanation.

The virtues of the Society of Friends are the virtues of liberalism: they cultivate honest discourse and right action, by taking care not to engage in practices that destroy individual discernment. The failings of the Society of Friends are the failings of liberalism: they do not seem to have the organizational capacity to recognize predatory systems and construct alternatives.

Fundamentally, Quaker protocols seem like a good start, but more articulated structures are necessary, especially more closed systems of production.

This post reflects a lot of thought, but there's a lot of speculation which I hope I've managed to mark as such. I'm optimizing for clearly communicating my present state in the hopes of furthering dialogue, not saying things that are maximally defensible; I haven't worked out the relevant models in extreme detail. That said, I don't think I'm misreporting any facts, and corrections on any level are welcome.

Some reasons to respect the Society of Friends

- Liberalism is nice, and the Quakers instilled it in America.
- They pioneered the radical practice of personal integrity.
- Their social technology is designed to avoid overriding individual conscience and judgment, thus preserving information that is typically destroyed by more common systems oriented around momentum or dominance.
- They don’t advertise much.

Proto-liberals

The Quakers first came to my attention when Scott Alexander of Slate Star Codex wrote about them. His review of Albion’s Seed describes them as proto-liberals with an outsized effect on the United States of America, basically winning over the culture to their ideals:

> Fischer warns against the temptation to think of the Quakers as normal modern people, but he has to warn us precisely because it’s so tempting. Where the Puritans seem like a dystopian caricature of virtue and the Cavaliers like a dystopian caricature of vice, the Quakers just seem ordinary. […]
>
> George Fox […] believed people were basically good and had an Inner Light that connected them directly to God without a need for priesthood, ritual, Bible study, or self-denial; mostly people just needed to listen to their consciences and be nice.

> Since everyone was equal before God, there was no point in holding up distinctions between lords and commoners: Quakers would just address everybody as “Friend”. And since the Quakers were among the most persecuted sects at the time, they developed an insistence on tolerance and freedom of religion which (unlike the Puritans) they stuck to even when shifting fortunes put them on top. They believed in pacificism, equality of the sexes, racial harmony, and a bunch of other things which seem pretty hippy-ish even today let alone in 1650.
>
> […] The Pennsylvanian leadership on abolitionism, penal reform, the death penalty, and so on all happened after the colony was officially no longer Quaker-dominated.
>
> And it’s hard not to see Quaker influence on the ideas of the modern US – which was after all founded in Philadelphia. In the middle of the Puritans demanding strict obedience to their dystopian hive society and the Cavaliers demanding everybody bow down to a transplanted nobility, the Pennsylvanians – who became the thought leaders of the Mid-Atlantic region including to a limited degree New York City – were pretty normal […] the Quakers really stand out in terms of freedom of religion, freedom of thought, checks and balances, and the idea of universal equality.

I like a lot of the opportunities that modern liberal society affords me. Some large amount of this good is attributable to the Quakers. This suggests that if I want more good things in this vein, I should check out what the Quakers are up to.

Personal integrity as a radical practice

I’ve come around to the point of view that personal integrity is not something I can expect from my environment by default. There are many social forces that corrode it, and it is only sustainable if conceived of as a radical practice.

The Quakers got there first. The Quaker insistence on not lying was wide-ranging. It included apparently little but socially costly things, like refusing to sign letters with the traditional “I remain your most obedient servant” unless they in fact remained the person’s most obedient servant, which generally they did not. It included more clearly materially costly things, like refusing to quote unrealistically high opening prices for the sake of bargaining, even when this was the predominant custom, preferring to quote the price they expected to charge.

Society has moved substantially towards the Quaker way on both of those practices, though honest pricing is still not reliably supported by market forces - revealed preferences are for slot machines. To do better takes something like a religion - a shared understanding that you’re not doing the profit-maximizing thing but the virtuous thing, and don’t expect that the outside world will always give you local incentives to do it.

Nonviolent social technology

Quaker worship values reflective silence. Quaker decisionmaking also centers empowering individuals to discern the right for themselves. "Clearness committees" provide guidance to Friends making difficult decisions, not through advice or admonition, but by asking questions to help the decider know the right with their own conscience. Quaker groups tend to favor decisionmaking models in which if even one

member continues to object, they either continue discussing the issue, or let it rest until later.

This stands in marked contrast to the predominant modes of group coordination, which focus on momentum or hierarchical control. Both those modes treat dissent as noise, to be eliminated. Quaker social technology treats it as signal, to be processed.

You might think that this would lead to total paralysis. But the question of racial slavery in the US is an instructive example. The Quakers did not all come to the right answer immediately. But they kept talking about it, and the argument "you wouldn't like it if someone did that to you" was sufficiently persuasive that (according to Scott's summary) during the 17th and 18th centuries slave ownership among the wealthy declined from 70% to 10%.

For a while I complained that our society's default understanding of friendship - and informal social relations in general - was built around momentum, and that everything else is unfairly rounded off to adversarial control. I yearned for a conception of trust that was based on shared discernment of the good for each other. It seems too right to be mere chance that when the group promoting the key virtues I thought necessary for true friendship chose a name for itself, that name was the Society of Friends.

Humble marketing

I've written a lot lately about the epistemically corrosive effects of marketing culture, and the coordination advantages of keeping a low profile under those circumstances. Quakers don't advertise much, they just seem to keep on doing sensible good things.

I am writing this from a cabin at a Quaker retreat. I wanted to take a couple weeks alone to reflect on my life and strategy, away from the pull and rhythm of any local social scene. It was surprisingly difficult to find a plain cabin that fit my needs.

One thing I tried was using modern search methods. AirBnB, Craigslist, Google, VRBO websites. Nearly every listing for a cabin or cottage was oriented towards vacationers' enjoyment, and outfitted and priced for a luxurious consumer experience. The ones that were not luxury cabins were for people experienced at camping or otherwise roughing it (e.g. people who know what to do without running water, which I do not). Nearly every offer of a silent retreat was for something like a managed experience with meditation practice, which I have found valuable enough to recommend, but which is not what I needed at this time. What I needed was a quiet place to stand upright, at a high vantage point, and survey the territory. With just the material comforts that I would be distracted without. Perhaps unsurprisingly, none of the vast profit-seeking marketing apparatus people use to find information was very helpful in finding a place to recover from its effects.

The other thing I tried was the old-fashioned, local-scale method: asking friends. I asked on my blog. I talked to people about it and got suggestions. This worked somewhat better. One kind friend offered his family's cabin, in New Hampshire. Another pointed me to a modern monastery that let out cabins, in Vermont, for a price comparable to the cheapest AirBnB options. Then, my mother mentioned my search to a friend, who told her about a Quaker retreat center in Massachusetts. I looked it up, and there were two similar retreat centers within driving distance of Berkeley (where I have been living), one of which had a cabin available. As far as I can tell, the Quakers operate these retreat centers more or less at cost, as a public service,

because they believe that people ought to have a place to go - groups as well as individuals - to reflect and discern the right path for themselves.

It's really remarkable how often, at how many different points in their history, they've been doing the exact most reasonable thing.

The best defense is strategy

So, if the Quakers are doing all these great things, why am I not one? A few years ago, my objection would have been that they are not focused on considerations of scope, and I'd have expected something like effective altruism to do much better. I no longer think that, because it seems to me like effective altruism in its current form is not epistemically sustainable. Local solutions, scaled organically at a rate compatible with human verification of results, have a significant advantage there.

My objections have more to do with the information-processing limitations of a totally nonhierarchical network that relies on peer-to-peer transfer of information. In particular, this on its own is not sufficient to avoid some systemic traps:

- Quaker coordination methods are inadequately defended against arbitrage.
- It takes a village to sustain life.
- You cannot serve two masters.

Arbitrageur defeats Quaker

The emphasis on doing good according to personal discernment, as an expression of personal conscience, rather than building the most scalable goodness marketing machine possible in order to maximize your impact, provides some resistance to the temptation to distort the truth to something more appealing. But it leaves you vulnerable to two types of arbitrage:

- If your environment does not have similarly good epistemic defenses, then you will still be the consumer of this kind of marketing, in the absence of a systemic plan to obtain accurate intelligence. I wrote about this in the humility argument for honesty, so I will not repeat myself here.
- If you reliably respond to local needs with outward-oriented service, you become an exploitable resource by more global strategies that may not share your values. I wrote about this in against neglectedness considerations, but feel that this needs more exposition.

Socially responsible investing vs vice funds

A simple example of the second kind of arbitrage is socially responsible investing. Some mutual funds avoid investing in harmful businesses businesses, such as arms dealers, tobacco companies, and casinos. The direct effect of this is to reduce demand for stock and debt in such companies, thus reducing the stock price and implicitly increasing their cost of capital. But if some businesses are systemically underpriced, this creates an arbitrage opportunity, to capture above-normal economic returns by investing in them. And in fact there are "vice" funds that tend to slightly outperform the market by doing exactly that.

This arbitrage does not appear to have completely negated the effects of socially responsible investing, but a substantial effect of this strategy is still to transfer money from people following a virtuous abstention policy, to those following an amoral one.

Volunteer work vs administrative efficiency

But what of something more local, like volunteer work? Suppose, for instance, that there is an epidemic. The legitimate authorities' responders, funded by taxes and wealthy prestigious foundations, are stretched thin, and you confirm that volunteers are needed, or identify an underserved area. You volunteer to tend to the sick and quarantined, and talk openly about this with friends, who then decide whether this is a thing that they ought to do.

The nice thing about this strategy is that you can be fairly sure that you are doing some local good. You can see for yourself that people are ill and suffering, and if you take care of yourself well enough to avoid spreading the epidemic, you can be reasonably sure you are helping locally. Since you are not using content-neutral persuasion tactics to mobilize large groups with unknown opportunity cost, you avoid imposing hidden costs globally. So far, so good.

What will the authorities' response be? The authorities initially understood that there was some chance of an epidemic, and made allowances for some systemic capacity to mitigate it. Their calculations took into account the costs of setting aside these resources, rather than doing something else with them.

If you are doing work that the authorities did not expect, then when they observe better than expected outcomes, they will factor this into their future plans, and reallocate resources away from epidemic preparedness, relative to the scenario where you did not participate.

If these authorities are benevolent - if they are optimizing for a metric that reflects your values well enough - then even if your net effect was mostly not to reduce death and suffering due to epidemics (because that was arbitraged away), you are still doing good, because you are freeing up resources to do other things, elsewhere, where you cannot personally verify opportunities to do good.

But it is far from guaranteed that the authorities are benevolent. If, on the other hand, their strategy is to spend as little as they can get away with on social services, in order to loot as much as possible from the system, then you have redistributed resources from yourself to them. This is one form of what economists call moral hazard.

Philanthropy vs moral hazard

Moral hazard is not limited to domains like business or government where the adversarial component is obvious; even in philanthropy, a field you might imagine characterized by an exceptionally high level of benevolence and value-alignment among donors, experts believe this problem exists. An instructive example is GiveWell, a charity evaluator which has consistently advised a foundation with more money than it knows what to do with to avoid fully funding GiveWell's top-recommended charities, in order to avoid this sort of moral hazard.

Fair trade vs business

When you are relating to an open system, in which you are a price-taker, engaging in unembedded transactions involving unknown agents with unknown agendas and strategies, you should expect everything you do to be arbitraged against. This is fine when you are trying to get something for yourself; if you buy a coffee at a cafe, you can for the most part personally verify that you have received a satisfactory coffee, and the value you receive does not really depend much on the hidden ways the cafe might respond to the incentive. Arbitrage is good, because it means that you are not creating local coffee scarcity when you buy your coffee; instead, you are sending a price signal that causes the global financial system to reallocate production very slightly towards coffee.

But what if you are worried about the negative externalities of your actions? What if you are worried that the coffee industry is extractive, in a way that harms some group of people with little economic leverage? You might try buying fair trade coffee. In effect, the fair trade label is an implicit guarantee about the net effects of your actions on the global economy. The obvious arbitrage opportunity here is to sell specious guarantees. Starbucks cannot sustainably mislead you about whether the cup they sell you has any coffee in it, but your ability to verify the desirable effects of fair trade claims is much weaker.

In praise of closed systems

Arbitrage is an inherent vulnerability of the outward-oriented, service model of right action. If you are looking to create value, you should favor closed systems where you (or trusted processes) can validate the accounting, observing the inputs and outputs, so that you can be sure that you end up with something more valuable than you started with. Accounting, not arbitrage.

Intentional communities and local production are examples of this. Building local creative and reflective capacity in ways that the official system is not optimizing for is a plausible way to create lasting value that does not immediately get arbitraged away.

This implies that some parts of the Quaker strategy are good for long-term value creation. In particular, the Friends meetings themselves, and the cultivation of individual discernment, are direct capacity-building exercises. I am, after all, writing this from a cabin at a Quaker retreat. The retreat center is obviously a good thing to exist, especially since it’s not marketed as an arbitrage opportunity for vacationers looking to get a better deal. There is even a local intentional community here. I am not completely clear on the details, but it seems as though some people live on site to maintain the retreat center.

These are really obviously the “good guys”; they’re supporting the development of vitally needed steering capacity. But this development is not enough; more research is needed.

It takes a village to sustain life

If things go well, this will not be the last generation of humans. This means that for long-term good outcomes, we need to bring up future generations. Unfortunately, there are many systemic forces working to make this difficult.

People who participate more in the abusive Western educational system tend to have fewer children.

Suburbanization makes it costly to raise children humanely; parents are forced to choose between sending their kids off to a designated abuse facility, or designating at least one parent to be a full-time caretaker. This work cannot be shared among communities to realize economies of scale, because most adults are busy far away at work, and in any event you can't let your kids run around freely because nearly every house abuts an active road with deadly automobile traffic.

[ETA: As Zvi points out, it is also effectively illegal in major US metropolitan areas to let your kids roam freely even when they are old enough that it is otherwise safe for them to do so.]

The net effect of all this is to make child-rearing an expensive consumption good, instead of an important part of the productive activities of life.

Intergenerational communities such as local religious groups often help mitigate this problem, but for the most part the Friends I have talked to did not seem to consider it an urgent priority, or be organizing their churches to fix this problem. Instead, they pay the costs if they can, and focus on helping the individuals most in need. This is probably related to the broader Christian orientation towards service rather than production.

I am happy to engage in further discourse about this with anyone who is interested, but this isn't something that can in principle be solved, I think, by incremental individual progress like eschewing leverage and manumitting your slaves. It requires coordinated action, and therefore group structures that can take decisive action even against local incentive gradients, with some amount of centralized responsibility, oriented around *group information-processing* rather than oriented merely around not destroying individual information-processing.

We need to learn how to *be free and build infrastructure*, or we will live in infrastructure built by and for an unfree world.

You cannot serve two masters

Religious minorities - so, in the major cities of the US, basically any religion aside from cosmopolitan secular liberalism - tend have lots of experience having to stand up for their practice at the expense of exclusion from communal events, institutional approval, and sometimes even livelihood. I have personal experience with this; Jews who insist on not working on the Sabbath or Holidays, for instance, frequently find this to be a source of friction with employers. As a child, I had to miss communal school events for this reason. In the Rationalist community, the upcoming autumn equinox celebration was scheduled to coincide with Yom Kippur.

Some employers talk of work/life balance, but ultimately you can only serve one master, and the relation most people have with their employers is one of a *servant in their household*. To pick a recent example, in this Twitter thread, a prominent journalist takes it as too obvious to be worth stating as an explicit premise, that the CEO of a major corporation can move some huge number of people to an arbitrary place to influence political decisions via "democracy." This looks very much not like people being free to me.

A moral community that doesn't organize around in-community production will quickly run up against the problem that you can have only one central organizing force in

your life. Either people will flake on their employers and be fired, or flake on community obligations and nothing will get done, or both.

A necessary part of any *viable alternative* to the world of marketing, is a community where people either own their own home & business, or depend for their livelihood on other community members who are committed to shared values.

I see some intentional communities at the extremes of Quaker life, but for the most part it looks like people are basically participating in the modern liberal world, with infrastructure unsuited to independence and human flourishing.

A very incomplete survey of other alternatives

I'm aware of a few other groups, in the West, that are doing, or used to be doing, *parts* of what is needed.

The Quakers are one of the four founding English cultures described in *Albion's Seed*. Another are the Puritans. They seem promising in a bunch of ways. They managed to pull off an integrated culture of shared production and communal norms, organized around communities and familial households that worked small landholdings together. They managed to have commerce, but also managed the moral leadership to at least occasionally slap down profit-maximization wireheading. They don't seem to have had much fun, which is unfortunate, but despite their heavy social control, they made sure to make room for private love and happiness.

The most obvious problem with the Puritans is that they don't exist anymore, because they weren't very good at making following generations enthusiastic Puritans. This is probably related to the ways in which they did not have much fun. The next most obvious problem, from my perspective, is that while they cared a lot about improving their land and lot, they wanted to do this in traditional ways; their resistance to innovations might have prevented them from advancing things like radical life extension and space travel, both things that are necessary if we want the good long-run future.

Jews fall into a few different interesting categories.

Liberal Jews, even Orthodox ones, primarily share community norms with each other, separating it from economic production and for the most part childrearing, which they do the conventional way. They face constant social and economic pressure to buckle, and often find themselves making awkward compromises. I've seen this work well for upper-middle-class Jews, but there's a reason why people are drifting away from the practice. It's expensive, in a world not designed to interface with it well. If your production is unrelated to your religion, then your religion becomes a *consumption good*, and the Sabbath is competing with Hollywood. Who's gonna appeal more to your kids? Liberal Judaism fails to make childrearing easy, though community does at least a little to make it easier. Liberal Jewish schooling and intellectual culture does seem to have preserved some sort of intellectual capacity above baseline, which contributes to value-aligned steering power as well as the sort of thing that wins Nobel prizes. Some of this might be genetic rather than cultural, of course, and I don't know how much.

Haredi Jews (the ones who wear lots of black) often work in community businesses, and have community infrastructure. Economic and child production and religion seem at least somewhat integrated. Why don't I join *them*? Well, I'd have to agree to a

bunch of things that I think are not true, and it doesn't seem like they're very *fast* at producing material progress, in part because they waste a lot of time pretending to be curious about things. They also seem to often have institutional child sexual abuse problems, like the Catholics, though it's not clear to me whether this is actually above the base rate of child molestation by authority figures, whether we simply find out about it more when distinctive minority-culture institutions have that problem, or something else is going on. But I would want to do much, much better, on that and many other fronts.

I also don't know whether there's something substantially more interesting going on in the yeshiva communities organized around schools that are the intellectual heirs of the Vilna Ga'on, vs Chassidic communities organized around a charismatic Rebbe.

I expect many religions have similar tradeoffs. Amish seem like they're in a similar position to Haredi Jews, with if anything much more local production, but less oriented towards scholarship, and they forgo many of the benefits of industrial technology. This might be fine for short-term quality of life (most Amish come back from Rumspringa, after all), but not a thing that can inherit the stars. Orson Scott Card's writing about being a Mormon outside traditionally Mormon areas like Utah is recognizably similar to the way committed liberal Jewish parents write about Judaism.

Academic communities seem to take a different horn of the liberal trilemma. They focus on the integrity of their *intellectual production* - which if not quite an economic production relationship, is at least an economic *service* relationship with the universities - but child-raising and personal ethics are regarded for the most part as a private matter. Colleges have clearly gotten worse, but part of this is that they have to work with the students they get, who seem to be getting worse year after year. (See this retrospective on Bloom's *The Closing of the American Mind* for a decent argument that scholarship is doomed without communities that bring up children to be good potential scholars.) Academics are neither in control of their own livelihoods (as communities they depend on outside funders) nor their other necessary inputs (since they don't regulate the upbringing of the students they teach, and are often even at the mercy of administrators for admissions).

My mostly-uninformed sense of hippie communes is that they seem to have the hang of living well, but don't do engineering to make progress or trade much or compete with with the industrial economy. This makes them much like the Amish, though perhaps they have more fun.

Anarchists are stereotyped as oppositional rather than focusing on building new alternatives, but I don't really know. My uninformed impression is that they are poor and unhappy, but again I really don't know.

And then there's Burning Man. Burning Man only lasts two weeks. That's not enough time to raise a new generation. It also has a taboo against producing artifacts of lasting value - you leave no trace. The main nice thing about it - by reputation, I haven't gone - is that it brings engineers and hippies together *at all*. See Kristina Miller's piece on Burning Man for a more detailed treatment.

Basically all these groups seem worth learning more about, now that I have eyes to see them with.

Of course, I might be wrong about the Quakers. They might actually have a lot going on that I haven't found out about yet, that answers all my objections. I could easily

not have heard of this - they don’t advertise much, after all. If so, I’d be delighted to learn about that.

Cross-posted at my personal blog.

Metamathematics and Probability

This is a linkpost for http://alexmennen.com/index.php/2017/09/22/metamathematics-and-probability/

Wikipedia pageviews: still in decline

In March 2015, I wrote about a decline in Wikipedia desktop pageviews over the last few years (and posted a short version to LessWrong). With a lot of help from Issa Rice over the last year, and a lot more quality data, I've revisited the claims of that post.

This post provides a high-level summary of my takeaways. If enough people express interest in the comments, I intend to write up in more detail on the aspects that people express interest in. If I do a more detailed writeup, it will probably be in the latter half of 2018, giving enough additional data to evaluate how well the decline hypothesis holds up.

Here are the top-level conclusions.

1. Have English desktop Wikipedia pageviews (i.e., pageviews of Wikipedia pages from desktop devices) actually declined?
   Short answer: Yes, they have declined by over 50% since the peak between late 2012 and late 2013. Some supposedly timeless page types have declined by up to 75-80%. The effect of per-page decline is partly cancelled by increase in number of pages.
   If I do a longer post, I'll compare the time periods September to November 2012 against September to November 2017, and April to June 2013 against April to June 2018. Both are three-month periods, with equal representation of all days of week, the same time of the year, and with a separation of five years.

2. Why have English desktop pageviews declined?
   Short answer: Substitution to mobile could explain between 10 and 40 percentage points of the desktop decline. I personally gravitate to the lower end of the estimate range.
   Inclusion/exclusion of non-human traffic could explain between 5 and 20 percentage points of the decline.
   Switch to HTTPS and the block of Wikipedia in China explain a sharp mid-2015 decline, but use of Chinese Wikipedia (which should have been most affected) has recovered, and I expect the long-term effect to be close to zero. At most, it is 5 percentage points.
   The residual decline is between 0 and 20 percentage points, which, after rebasing, is between 0 and 40% for desktop. Two leading candidates to explain the residual are *increased reliance on social media* and *search engine algorithm changes*.

3. Have total (desktop + mobile) human English Wikipedia pageviews declined? Why?
   Short answer: Total (desktop + mobile) human pageviews likely peaked around late 2013, and have declined by about 20% since then. Per-page pageviews have gone down significantly more for the page types that saw the biggest desktop declines. Effect of per-page decline is partly cancelled by increase in number of pages.
   Candidate explanations are the same as for (2): *increased reliance on social media* and *search engine algorithm changes*.

4. Is there a compensating increase in other language Wikipedias?
   Short answer: No. In fact, other top language Wikipedias (German, Russian,

Spanish, Japanese, French) have a broadly similar decline trend as the English Wikipedia, both overall and per-page.
Some minor language Wikipedias saw a huge proportional increase but not enough to compensate for the English Wikipedia decline. For instance, monthly Hindi Wikipedia mobile web pageviews exploded from about 1 million in early 2013 to over 30 million in 2017, which is peanuts compared to 3-4 billion monthly English desktop and English mobile web Wikipedia pageviews.
The lowest-traffic language Wikipedias saw a huge proportional decline in desktop and mobile web traffic in 2015, which is explained by bot filtering being activated.

5. Do people subjectively feel they are using Wikipedia less? How do we square their subjective impressions with the statistics?
   People generally perceive either no change in use or say they don't use Wikipedia at all.
   But in a head-to-head comparison of "use more now" versus "use less now", the former wins.

Why might this be an interesting thing to study?

Wikipedia pageview data is one of the most comprehensive and granular open datasets covering a wide variety of areas of interest, so they provide a useful way to understand both people's *relative interest in different topics*, and the *trends in individual topics as well as the Internet as a whole*. Specifically:

1. If you're interested in how interest in specific topics has evolved over time, or if you're interested in how people's Internet use has changed over time, Wikipedia pageviews are a useful part of your toolkit, just like Google Trends. Having a good sense of the general trends in Wikipedia pageviews allows you to better "normalize" for these trends and give more context to the numbers you see.

2. If you're interested in the overall growth (or decline!) of the Internet, Wikipedia, as one of the top sites on the Internet, and one that does not engage in a lot of advertising and view optimization, offers some insight.

3. One of the hypotheses that might explain part of the decline, namely *increased reliance on social media*, is of particular interest to rationalists and LessWrong. LessWrong pageviews also peaked at roughly the same time as Wikipedia pageviews, and social media (particularly Facebook) has been implicated in the decline of LessWrong (see the comments here).

So, what do you think? How interesting do you find this topic? What parts are you skeptical of? What parts are you most interested in seeing explored or justified more rigorously?

PS: If you're curious what a more detailed report might look like, check out the draft Issa and I worked on last year. All responsibility for errors, both in the draft and in this teaser post, is mine. You can also check out the timeline of Wikimedia analytics to understand changes relevant to interpreting analytics.

Tests make creating AI hard

In my post on incentive structures I gave an potted summary of how to do incentive structures better when what you are trying to achieve is ill defined,

> Improve Models using Measures, use the Model to update Targets.

I would add,

> Try to hit Targets. Avoid Tests.

In this post I will recap what I mean by the above, give an abridged history of the field of AI and how it has failed to follow this maxim well and what following it might have looked like.

If we want to create safe AI we need good incentive structures for the people researching it, we need to get good at this.

Recap

A Model is your causal view of the thing you are trying to achieve. A Measure is something you can apply to your system to let you know if you are going in the right direction. Targets are things you are trying to do in the world. A Test in this formalism is something that is a Measure and Target all in one, you don't need a Model, if you do well at the Test you are doing well.

One of the key differences between a Measure and Test, if you do better than a Model predicts on a Measure you should change your Model (and maybe change your Target). If you do better than you expect on a Test, there is no need to change anything.

Measure 1: The Turing test

The most famous measure in AI is the Turing test.

> [it] is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. Turing proposed that a human evaluator would judge natural language conversations between a human and a machine designed to generate human-like responses. The evaluator would be aware that one of the two partners in conversation is a machine, and all participants would be separated from one another.

From this we got the Loebner prize. This in turn has led to profusion of chat bots like this one. It might entertain the judges and give them a moments pause to try and figure out whether it is a human or not, but you can't get much real work out of it. It can't do maths, run a business or teach kids. It is a product of taking the Measure as a Test.

However it is still a good Measure: if something really passed it we would expect it to be intelligent.

Measure 2: Image net

This is not supposed to be a test of full intelligence, but how well an algorithm detects object in images . While it is creating algorithms that do better and better at this all the time, it seems unlikely that it is getting close to how humans do object recognition.

For example the algorithms aren't being designed to take hints about what is in an image to help it locate the object, because that is not what they are being tested for. Why might you want to do so? Because there are lots of good examples humans being able to process these hints. This seems like a useful thing to do. You can experience the power of "hint taking" first hand, if you read this slatestarcodex article.

If your Measure and your Target is static, you are just trying to make a slightly better algorithm that does better at these Tests. You are not going to change your Model of image recognition to be able to incorporate other types of data, like "hints".

Breaking the Test cycle

So we are in general teaching our system to the tests. And when the the Tests are iterated upon (because they don't get what people actually want), the people trying to iterate the Tests get accused of moving the goal posts. There is even a name for it, the AI effect.

If we are to get safe general AI or IA we need to break this cycle.
We need a Model of what intelligence is and iterate on that, so that we can get different Targets. Not naively try to meet the Tests better or add more and more Tests.

I obviously think separating Measures and Metrics will help us with making safe AI, so how can we

> Improve Models using the Measure, use the Model to update Targets. Hit Targets, Avoid Tests.

The following an illustrative alternative history about the development AI could've gone gone if we had better separated our Targets from our Measures. It will be described by a series of rounds. Each round will have an assumed Model, a Target to try and create due to that model, the Measure (in this case the Turing test, but you could have more measures) and a result of having performed that measurement. There will also be a failure mode to avoid in the round.

Round 1:

Model: Let us start with a model of human conversation as a fixed input/output mapping, for simplicity's sake.

Target: Create a mapping of input to output that is nice to talk to.
Measure: Use the turing test as a measure.

Result: This isn't very good to talk to. It is the same every time. Humans aren't the same every time. Update the model of an intelligence so that it isn't seen as a fixed mapping.

Failure mode: Create ever more elaborate mappings from input to output, include previous parts of the conversation in the input.

Round 2

Update Model: Humans seem to learn over time. So let us assume an intelligence also learns.

Target: Create a machine learning system that tries to find a mapping from input to output from data. Provide the system with the data from previous conversations and how well the judges like those conversations. So that it can update it's mapping from input to output.

Measure: Use the turing test as a measure.

Result: It was pleasant to talk to but one judge tried to teach it simple mathematics (adding two numbers together) and it failed. There is no way our current model of an intelligence could learn mathematics in a single conversation.

Failure Mode: Throw more and more data and processing power at it, creating ever more complex mappings

Round 3

Update Model: Humans seem to be able to treat natural language as programs to be interpreted and compiled. They learn languages by this using large amounts of data, but they also use language to help learn other bits of language.

Target: Create some system that can discern good programs from bad programs, then put programs in it that try to compile human language into novel programs. Also have programs that search for novel patterns in the input language and tries to hook them up to code generation.

Measure: Use the turing test as a measure.

Result:??

Conclusion

I hope I have shown you how the maxim might be useful for thinking about incentive structures for solving a complex problem that we don't quite understand what we are aiming for. While this is just my initial Model of what we should do for incentive structures for human researchers, it is very important to get right (as it could be used inside AI as well as how to build it).
So to move forward on the intelligence problem we need to create the best model of intelligence we can, so we can find targets. I think I have one (Round 3), but I would I love to get more input into it.

It would also be interesting to think about when having a Model would be a bad idea and it would better to just use Tests. There are probably some circumstances.

An incentive structure that might not suck too much.

You want to make things better or live in a world which makes things better. But how do you go about actually doing that? There is truth to Drucker's maxim

> "If you can't measure it, you can't improve it."

But you have also heard of Goodheart's law.

> "When a measure becomes a target, it ceases to be a good measure."

But you have measures. How can you use them without them becoming a target?

Delegation

If you are delegating work this becomes even trickier. You can't give the people the you are delegating the measures you use (unless you trust them to use them properly and not use them as targets). So you give people rules to follow or goals to meet that aren't the measures.

Then evaluate them on how well they follow the rules or goals (not on how well they meet the measure), and iterate on the rules or goals. If you have people that follow the rules and achieve the goals *AND* you iterate on those things you can actually affect change in the world. If you don't iterate you'll just end up optimising whatever the first set of rules points at, rather than the thing you actually want to acheive.

You also probably want to give people some slack with the rules/goals, so that they have spare energy to look at the world and figure out what they think is best. If people are run ragged trying to meet a goal to survive, all other considerations fall by the way side.

Fixing the goals

During the iteration of the rules, how do you avoid Goodheart's law yourself? You want people to lead good happy lives, but don't want to end up secretly giving people drugs to make them happy, because you are short-circuiting things. You also don't want to kill everyone to reduce long term suffering.

So instead you build yourself a model of what Good looks like. This model is important it allows you to decouple your measure from your target. An example might be, "It is Good for People to be wealthy as it allows them to do more things". You use your model to generate a target, in this case "make people wealthier". Then you alter the rules and goals to hit that target.

What happens if your model is wrong? Let us say some people are becoming unhappier as they become wealthier, due to pollution causing health issues.

This is where the measure comes in. You then use your measures to see if your model is correct. If people aren't becoming happier, less stressed, healthier etc as they become wealthier, you change and update your model. In this case they aren't, so you

improve your model, find a new targets and therefore give new goals and rules to the people you delegate to.

Any anger at the poor performance related to your measure should not be taken out on the people you have delegated to (unless they didn't do what you said), it should be taken out on your model of the world that thought it was idea to tell people to do something.

Also you can improve your model with small scale studies and trying to understand the inner workings of humans. This gives you a quicker feedback loop than changing society and seeing what happens.

This is a description of roughly where we are as a society currently, although we suck at the last section updating our models and changing our targets. We tend get stuck on the first set of targets we find, those of GDP and IQ, publishing 'high impact' papers or reducing waiting times at doctors. We don't use measures to say, hey somethings wrong, let us change things. The best we have is Democracy but that is a very blunt instrument and has its own incentive structure problems.

I Can Tolerate Anything Except The Outgroup

[Content warning: Politics, religion, social justice, spoilers for "The Secret of Father Brown". This isn't especially original to me and I don't claim anything more than to be explaining and rewording things I have heard from a bunch of other people. Unapologetically America-centric because I'm not informed enough to make it otherwise. Try to keep this off Reddit and other similar sorts of things.]

I.

In Chesterton's *The Secret of Father Brown*

, a beloved nobleman who murdered his good-for-nothing brother in a duel thirty years ago returns to his hometown wracked by guilt. All the townspeople want to forgive him immediately, and they mock the titular priest for only being willing to give a measured forgiveness conditional on penance and self-reflection. They lecture the priest on the virtues of charity and compassion.

Later, it comes out that the beloved nobleman did *not* in fact kill his good-for-nothing brother. The good-for-nothing brother killed the beloved nobleman (and stole his identity). *Now* the townspeople want to see him lynched or burned alive, and it is only the priest who – consistently – offers a measured forgiveness conditional on penance and self-reflection.

The priest tells them:

> It seems to me that you only pardon the sins that you don't really think sinful. You only forgive criminals when they commit what you don't regard as crimes, but rather as conventions. You forgive a conventional duel just as you forgive a conventional divorce. You forgive because there isn't anything to be forgiven.

He further notes that this is why the townspeople can self-righteously consider themselves more compassionate and forgiving than he is. Actual forgiveness, the kind the priest needs to cultivate to forgive evildoers, is really really hard. The fake forgiveness the townspeople use to forgive the people they like is really easy, so they get to boast not only of their forgiving nature, but of how much nicer they are than those mean old priests who find forgiveness difficult and want penance along with it.

After some thought I agree with Chesterton's point. There are a lot of people who say "I forgive you" when they mean "No harm done", and a lot of people who say "That was unforgiveable" when they mean "That was genuinely really bad". Whether or not forgiveness is *right* is a complicated topic I do not want to get in here. But since forgiveness is generally considered a virtue, and one that many want credit for having, I think it's fair to say you only earn the right to call yourself 'forgiving' if you forgive things that genuinely hurt you.

To borrow Chesterton's example, if you think divorce is a-ok, then you don't get to "forgive" people their divorces, you merely ignore them. Someone who thinks divorce

is abhorrent can “forgive” divorce. *You* can forgive theft, or murder, or tax evasion, or something *you* find abhorrent.

I mean, from a utilitarian point of view, you are still doing the correct action of not giving people grief because they’re a divorcee. You can have all the Utility Points you want. All I’m saying is that if you “forgive” something you don’t care about, you don’t earn any Virtue Points.

(by way of illustration: a billionaire who gives $100 to charity gets as many Utility Points as an impoverished pensioner who donates the same amount, but the latter gets a lot more Virtue Points)

Tolerance is also considered a virtue, but it suffers the same sort of dimished expectations forgiveness does.

The Emperor summons before him Bodhidharma and asks: “Master, I have been tolerant of innumerable gays, lesbians, bisexuals, asexuals, blacks, Hispanics, Asians, transgender people, and Jews. How many Virtue Points have I earned for my meritorious deeds?”

Bodhidharma answers: “None at all”.

The Emperor, somewhat put out, demands to know why.

Bodhidharma asks: “Well, what do you think of gay people?”

The Emperor answers: “What do you think I am, some kind of homophobic bigot? Of course I have nothing against gay people!”

And Bodhidharma answers: “Thus do you gain no merit by tolerating them!”

II.

If I had to define “tolerance” it would be something like “respect and kindness toward members of an outgroup”.

And today we have an almost unprecedented situation.

We have a lot of people – like the Emperor – boasting of being able to tolerate everyone from every outgroup they can imagine, loving the outgroup, writing long paeans to how great the outgroup is, staying up at night fretting that somebody else might not like the outgroup enough.

This is really surprising. It’s a total reversal of everything we know about human psychology up to this point. No one did any genetic engineering. No one passed out weird glowing pills in the public schools. And yet suddenly we get an entire group of people who conspicuously promote and defend their outgroups, the outer the better.

What is going on here?

Let’s start by asking what exactly an outgroup is.

There’s a very boring sense in which, assuming the Emperor’s straight, gays are part of his “outgroup” ie a group that he is not a member of. But if the Emperor has curly hair, are straight-haired people part of his outgroup? If the Emperor’s name starts with the letter ‘A’, are people whose names start with the letter ‘B’ part of his outgroup?

Nah. I would differentiate between multiple different meanings of outgroup, where one is “a group you are not a part of” and the other is…something stronger.

I want to avoid a very easy trap, which is saying that outgroups are about how different you are, or how hostile you are. I don’t think that’s quite right.

Compare the Nazis to the German Jews and to the Japanese. The Nazis were very similar to the German Jews: they looked the same, spoke the same language, came from a similar culture. The Nazis were totally different from the Japanese: different race, different language, vast cultural gap. But the Nazis and Japanese mostly got along pretty well. Heck, the Nazis were actually moderately positively disposed to the *Chinese*, even when they were technically at war. Meanwhile, the conflict between the Nazis and the German Jews – some of whom didn’t even realize they were anything other than German until they checked their grandparents’ birth certificate – is the stuff of history and nightmares. Any theory of outgroupishness that naively assumes the Nazis’ natural outgroup is Japanese or Chinese people will be totally inadequate.

And this isn’t a weird exception. Freud spoke of the narcissism of small differences, saying that “it is precisely communities with adjoining territories, and related to each other in other ways as well, who are engaged in constant feuds and ridiculing each other”. Nazis and German Jews. Northern Irish Protestants and Northern Irish Catholics. Hutus and Tutsis. South African whites and South African blacks. Israeli Jews and Israeli Arabs. Anyone in the former Yugoslavia and anyone else in the former Yugoslavia.

So what makes an outgroup? Proximity plus small differences. If you want to know who someone in former Yugoslavia hates, don’t look at the Indonesians or the Zulus or the Tibetans or anyone else distant and exotic. Find the Yugoslavian ethnicity that lives closely intermingled with them and is most conspicuously similar to them, and chances are you’ll find the one who they have eight hundred years of seething hatred toward.

What makes an unexpected in-group? The answer with Germans and Japanese is obvious – a strategic alliance. In fact, the World Wars forged a lot of unexpected temporary pseudo-friendships. A recent article from War Nerd points out that the British, after spending centuries subjugating and despising the Irish and Sikhs, suddenly needed Irish and Sikh soldiers for World Wars I and II respectively. “Crush them beneath our boots” quickly changed to fawning songs about how “there never was a coward where the shamrock grows” and endless paeans to Sikh military prowess.

Sure, scratch the paeans even a little bit and you find condescension as strong as ever. But eight hundred years of the British committing genocide against the Irish and considering them literally subhuman turned into smiles and songs about shamrocks once the Irish started looking like useful cannon fodder for a larger fight. And the Sikhs, dark-skinned people with turbans and beards who pretty much exemplify the European stereotype of “scary foreigner”, were lauded by everyone from the news media all the way up to Winston Churchill.

In other words, outgroups may be the people who look exactly like you, and scary foreigner types can become the in-group on a moment’s notice when it seems convenient.

III.

There are certain theories of dark matter where it barely interacts with the regular world *at all*, such that we could have a dark matter planet exactly co-incident with Earth and never know. Maybe dark matter people are walking all around us and through us, maybe my house is in the Times Square of a great dark matter city, maybe a few meters away from me a dark matter blogger is writing on his dark matter computer about how weird it would be if there was a light matter person he couldn't see right next to him.

This is sort of how I feel about conservatives.

I don't mean the sort of light-matter conservatives who go around complaining about Big Government and occasionally voting for Romney. I see those guys all the time. What I mean is – well, take creationists. According to Gallup polls, about 46% of Americans are creationists. Not just in the sense of believing God helped guide evolution. I mean they think evolution is a vile atheist lie and God created humans exactly as they exist right now. That's half the country.

And I don't have a *single one of those people* in my social circle. It's not because I'm deliberately avoiding them; I'm pretty live-and-let-live politically, I wouldn't ostracize someone just for some weird beliefs. And yet, even though I probably know about a hundred fifty people, I am pretty confident that not one of them is creationist. Odds of this happening by chance? 1/2^150 = 1/10^45 = approximately the chance of picking a particular atom if you are randomly selecting among all the atoms on Earth.

About forty percent of Americans want to ban gay marriage. I think if I *really* stretch it, maybe ten of my top hundred fifty friends might fall into this group. This is less astronomically unlikely; the odds are a mere one to one hundred quintillion against.

People like to talk about social bubbles, but that doesn't even begin to cover one hundred quintillion. The only metaphor that seems really appropriate is the bizarre dark matter world.

I live in a Republican congressional district in a state with a Republican governor. The conservatives are definitely out there. They drive on the same roads as I do, live in the same neighborhoods. But they might as well be made of dark matter. I never meet them.

To be fair, I spend a lot of my time inside on my computer. I'm browsing sites like Reddit.

Recently, there was a thread on Reddit asking – Redditors Against Gay Marriage, What Is Your Best Supporting Argument? A Reddit user who didn't understand how anybody could be against gay marriage honestly wanted to know how other people who *were* against it justified their position. He figured he might as well ask one of the largest sites on the Internet, with an estimated user base in the tens of millions.

It soon became clear that nobody there was actually against gay marriage.

There were a bunch of posts saying "I of course support gay marriage but here are some reasons some other people might be against it," a bunch of others saying "my argument against gay marriage is the government shouldn't be involved in the marriage business at all", and several more saying "why would you even ask this question, there's no possible good argument and you're wasting your time". About halfway through the thread someone started saying homosexuality was unnatural and I *thought* they were going to be the first one to actually answer the question, but at

the end they added “But it’s not my place to decide what is or isn’t natural, I’m still pro-gay marriage.”

In a thread with 10,401 comments, a thread *specifically* asking for people against gay marriage, I was eventually able to find *two* people who came out and opposed it, way near the bottom. Their posts started with “I know I’m going to be downvoted to hell for this…”

But I’m not only on Reddit. I also hang out on LW.

On last year’s survey, I found that of American LWers who identify with one of the two major political parties, 80% are Democrat and 20% Republican, which actually sounds pretty balanced compared to some of these other examples.

But it doesn’t last. Pretty much all of those “Republicans” are libertarians who consider the GOP the lesser of two evils. When allowed to choose “libertarian” as an alternative, only 4% of visitors continued to identify as conservative. But that’s still… some. Right?

When I broke the numbers down further, 3 percentage points of those are neoreactionaries, a bizarre sect that wants to be ruled by a king. Only *one percent* of LWers were normal everyday God-‘n-guns-but-not-George-III conservatives of the type that seem to make up about half of the United States.

It gets worse. My formative years were spent at a university which, if it was similar to other elite universities, had a faculty and a student body that skewed about 90-10 liberal to conservative – and we can bet that, like LW, even those few token conservatives are Mitt Romney types rather than God-n’-guns types. I get my news from vox.com, an Official Liberal Approved Site. Even when I go out to eat, it turns out my favorite restaurant, California Pizza Kitchen, is the most liberal restaurant in the United States.

I inhabit the same geographical area as *scores and scores* of conservatives. But without meaning to, I have created an *outrageously* strong bubble, a 10^45 bubble. Conservatives are all around me, yet I am about as likely to have a serious encounter with one as I am a Tibetan lama.

(Less likely, actually. One time a Tibetan lama came to my college and gave a really nice presentation, but if a conservative tried that, people would protest and it would be canceled.)

IV.

One day I realized that entirely by accident I was fulfilling *all* the Jewish stereotypes.

I’m nerdy, over-educated, good with words, good with money, weird sense of humor, don’t get outside much, I like deli sandwiches. And I’m a psychiatrist, which is about the most stereotypically Jewish profession short of maybe stand-up comedian or rabbi.

I’m not very religious. And I don’t go to synagogue. But *that’s* stereotypically Jewish too!

I bring this up because it would be a mistake to think “Well, a Jewish person is by definition someone who is born of a Jewish mother. Or I guess it sort of also means someone who follows the Mosaic Law and goes to synagogue. But I don’t care about

Scott's mother, and I know he doesn't go to synagogue, so I can't gain any useful information from knowing Scott is Jewish."

The defining factors of Judaism – Torah-reading, synagogue-following, mother-having – are the tip of a giant iceberg. Jews sometimes identify as a "tribe", and even if you don't attend synagogue, you're still a member of that tribe and people can still (in a statistical way) infer things about you by knowing your Jewish identity – like how likely they are to be psychiatrists.

The last section raised a question – if people rarely select their friends and associates and customers explicitly for politics, how do we end up with such intense political segregation?

Well, in the same way "going to synagogue" is merely the iceberg-tip of a Jewish tribe with many distinguishing characteristics, so "voting Republican" or "identifying as conservative" or "believing in creationism" is the iceberg-tip of a conservative tribe with many distinguishing characteristics.

A disproportionate number of my friends are Jewish, because I meet them at psychiatry conferences or something – we self-segregate not based on explicit religion but on implicit tribal characteristics. So in the same way, political tribes self-segregate to an impressive extent – a 1/10^45 extent, I will never tire of hammering in – based on their implicit tribal characteristics.

The people who are actually into this sort of thing sketch out a bunch of speculative tribes and subtribes, but to make it easier, let me stick with two and a half.

The Red Tribe is most classically typified by conservative political beliefs, strong evangelical religious beliefs, creationism, opposing gay marriage, owning guns, eating steak, drinking Coca-Cola, driving SUVs, watching lots of TV, enjoying American football, getting conspicuously upset about terrorists and commies, marrying early, divorcing early, shouting "USA IS NUMBER ONE!!!", and listening to country music.

The Blue Tribe is most classically typified by liberal political beliefs, vague agnosticism, supporting gay rights, thinking guns are barbaric, eating arugula, drinking fancy bottled water, driving Priuses, reading lots of books, being highly educated, mocking American football, feeling vaguely like they should like soccer but never really being able to get into it, getting conspicuously upset about sexists and bigots, marrying later, constantly pointing out how much more civilized European countries are than America, and listening to "everything except country".

(There is a partly-formed attempt to spin off a Grey Tribe typified by libertarian political beliefs, Dawkins-style atheism, vague annoyance that the question of gay rights even comes up, eating paleo, drinking Soylent, calling in rides on Uber, reading lots of blogs, calling American football "sportsball", getting conspicuously upset about the War on Drugs and the NSA, and listening to filk – but for our current purposes this is a distraction and they can safely be considered part of the Blue Tribe most of the time)

I think these "tribes" will turn out to be even stronger categories than politics. Harvard might skew 80-20 in terms of Democrats vs. Republicans, 90-10 in terms of liberals vs. conservatives, but maybe 99-1 in terms of Blues vs. Reds.

It's the many, many differences between these tribes that explain the strength of the filter bubble – which *have I mentioned* segregates people at a strength of 1/10^45?

Even in something as seemingly politically uncharged as going to California Pizza Kitchen or Sushi House for dinner, I'm restricting myself to the set of people who like cute artisanal pizzas or sophsticated foreign foods, which are classically Blue Tribe characteristics.

Are these tribes based on geography? Are they based on race, ethnic origin, religion, IQ, what TV channels you watched as a kid? I don't know.

Some of it is certainly genetic – estimates of the genetic contribution to political association range from 0.4 to 0.6. Heritability of one's attitudes toward gay rights range from 0.3 to 0.5, which hilariously is a little more heritable than homosexuality itself.

(for an interesting attempt to break these down into more rigorous concepts like "traditionalism", "authoritarianism", and "in-group favoritism" and find the genetic loading for each see here. For an attempt to trace the specific genes involved, which mostly turn out to be NMDA receptors, see here)

But I don't think it's just genetics. There's something else going on too. The word "class" seems like the closest analogue, but only if you use it in the sophisticated Paul Fussell *Guide Through the American Status System* way instead of the boring "another word for how much money you make" way.

For now we can just accept them as a brute fact – as multiple coexisting societies that might as well be made of dark matter for all of the interaction they have with one another – and move on.

V.

The worst reaction I've ever gotten to a blog post was when I wrote about the death of Osama bin Laden. I've written all sorts of stuff about race and gender and politics and whatever, but that was the worst.

I didn't come out and say I was happy he was dead. But some people interpreted it that way, and there followed a bunch of comments and emails and Facebook messages about how could I possibly be happy about the death of another human being, even if he was a bad person? Everyone, even Osama, is a human being, and we should never rejoice in the death of a fellow man. One commenter came out and said:

> I'm surprised at your reaction. As far as people I casually stalk on the internet (ie, LJ and Facebook), you are the first out of the "intelligent, reasoned and thoughtful" group to be uncomplicatedly happy about this development and not to be, say, disgusted at the reactions of the other 90% or so.

This commenter was right. Of the "intelligent, reasoned, and thoughtful" people I knew, the overwhelming emotion was conspicuous disgust that other people could be happy about his death. I hastily backtracked and said I wasn't happy per se, just surprised and relieved that all of this was finally behind us.

And I genuinely believed that day that I had found some unexpected good in people – that everyone I knew was so humane and compassionate that they were unable to rejoice even in the death of someone who hated them and everything they stood for.

Then a few years later, Margaret Thatcher died. And on my Facebook wall – made of these same "intelligent, reasoned, and thoughtful" people – the most common

response was to quote some portion of the song “Ding Dong, The Witch Is Dead”. Another popular response was to link the videos of British people spontaneously throwing parties in the street, with comments like “I wish I was there so I could join in”. From this exact same group of people, not a single expression of disgust or a “c’mon, guys, we’re all human beings here.”

I gently pointed this out at the time, and mostly got a bunch of “yeah, so what?”, combined with links to an article claiming that “the demand for respectful silence in the wake of a public figure’s death is not just misguided but dangerous”.

And that was when something clicked for me.

You can talk all you want about Islamophobia, but my friend’s “intelligent, reasoned, and thoughtful people” – her name for the Blue Tribe – can’t get together enough energy to really hate Osama, let alone Muslims in general. We understand that what he did was bad, but it didn’t anger us personally. When he died, we were able to very rationally apply our better nature and our Far Mode beliefs about how it’s never right to be happy about anyone else’s death.

On the other hand, that same group absolutely *loathed* Thatcher. Most of us (though not all) can agree, if the question is posed explicitly, that Osama was a worse person than Thatcher. But in terms of actual gut feeling? Osama provokes a snap judgment of “flawed human being”, Thatcher a snap judgment of “scum”.

I started this essay by pointing out that, despite what geographical and cultural distance would suggest, the Nazis’ outgroup was not the vastly different Japanese, but the almost-identical German Jews.

And my hypothesis, stated plainly, is that if you’re part of the Blue Tribe, then your outgroup isn’t al-Qaeda, or Muslims, or blacks, or gays, or transpeople, or Jews, or atheists – it’s the Red Tribe.

VI.

“But racism and sexism and cissexism and anti-Semitism are these giant all-encompassing social factors that verge upon being human universals! Surely you’re not arguing that mere *political* differences could ever come close to them!”

One of the ways we *know* that racism is a giant all-encompassing social factor is the Implicit Association Test. Psychologists ask subjects to quickly identify whether words or photos are members of certain gerrymandered categories, like “either a white person’s face or a positive emotion” or “either a black person’s face and a negative emotion”. Then they compare to a different set of gerrymandered categories, like “either a black person’s face or a positive emotion” or “either a white person’s face or a negative emotion.” If subjects have more trouble (as measured in latency time) connecting white people to negative things than they do white people to positive things, then they probably have subconscious positive associations with white people. You can try it yourself here.

Of course, what the test famously found was that even white people who claimed to have no racist attitudes at all usually had positive associations with white people and negative associations with black people on the test. There are very many claims and counterclaims about the precise meaning of this, but it ended up being a big part of the evidence in favor of the current consensus that all white people are at least a little racist.

Anyway, three months ago, someone finally had the bright idea of doing an Implicit Association Test with political parties, and they found that people's unconscious partisan biases were *half again as strong* as their unconscious racial biases (h/t Bloomberg. For example, if you are a white Democrat, your unconscious bias against blacks (as measured by something called a d-score) is 0.16, but your unconscious bias against Republicans will be 0.23. The Cohen's *d* for racial bias was 0.61, by the book a "moderate" effect size; for party it was 0.95, a "large" effect size.

Okay, fine, but we know race has *real world* consequences. Like, there have been several studies where people sent out a bunch of identical resumes except sometimes with a black person's photo and other times with a white person's photo, and it was noticed that employers were much more likely to invite the fictional white candidates for interviews. So just some stupid Implicit Association Test results can't compare to that, right?

Iyengar and Westwood also decided to do the resume test for parties. They asked subjects to decide which of several candidates should get a scholarship (subjects were told this was a genuine decision for the university the researchers were affiliated with). Some resumes had photos of black people, others of white people. And some students listed their experience in Young Democrats of America, others in Young Republicans of America.

Once again, discrimination on the basis of party was much stronger than discrimination on the basis of race. The size of the race effect for white people was only 56-44 (and in the reverse of the expected direction); the size of the party effect was about 80-20 for Democrats and 69-31 for Republicans.

If you want to see their third experiment, which applied *yet another* classic methodology used to detect racism and *once again* found partyism to be much stronger, you can read the paper.

I & W did an unusually thorough job, but this sort of thing isn't new or ground-breaking. People have been studying "belief congruence theory" – the idea that differences in beliefs are more important than demographic factors in forming in-groups and outgroups – for decades. As early as 1967, Smith et al were doing surveys all over the country and finding that people were more likely to accept friendships across racial lines than across beliefs; in the forty years since then, the observation has been replicated scores of times. Insko, Moe, and Nacoste's 2006 review Belief Congruence And Racial Discrimination concludes that:

> . The literature was judged supportive of a weak version of belief congruence theory which states that in those contexts in which social pressure is nonexistent or ineffective, belief is more important than race as a determinant of racial or ethnic discrimination. Evidence for a strong version of belief congruence theory (which states that in those contexts in which social pressure is nonexistent, or ineffective, belief is the only determinant of racial or ethnic discrimination) and was judged much more problematic.

One of the best-known examples of racism is the "Guess Who's Coming To Dinner" scenario where parents are scandalized about their child marrying someone of a different race. Pew has done some good work on this and found that only 23% of conservatives and 1% (!) of liberals admit they would be upset in this situation. But Pew *also* asked how parents would feel about their child marrying someone of a different *political party*. Now 30% of conservatives and 23% of liberals would get

upset. Average them out, and you go from 12% upsetness rate for race to 27% upsetness rate for party – more than double. Yeah, people do lie to pollsters, but a picture is starting to come together here.

(Harvard, by the way, is a tossup. There are more black students – 11.5% – than conservative students – 10% – but there are more conservative faculty than black faculty.)

Since people will delight in misinterpreting me here, let me overemphasize what I am *not* saying. I'm not saying people of either party have it "worse" than black people, or that partyism is more of a *problem* than racism, or any of a number of stupid things along those lines which I am sure I will nevertheless be accused of believing. Racism is worse than partyism because the two parties are at least kind of balanced in numbers and in resources, whereas the brunt of an entire country's racism falls on a few underprivileged people. I am saying that the *underlying attitudes that produce* partyism are stronger than the underlying attitudes that produce racism, with no necessary implications on their social effects.

But if we want to look at people's psychology and motivations, partyism and the particular variant of tribalism that it represents are going to be fertile ground.

VII.

Every election cycle like clockwork, conservatives accuse liberals of not being sufficiently pro-America. And every election cycle like clockwork, liberals give extremely unconvincing denials of this.

"It's not that we're, like, *against* America per se. It's just that…well, did you know Europe has much better health care than we do? And much lower crime rates? I mean, come on, how did they get so awesome? And we're just sitting here, can't even get the gay marriage thing sorted out, seriously, what's wrong with a country that can't… sorry, what were we talking about? Oh yeah, America. They're okay. Cesar Chavez was really neat. So were some other people outside the mainstream who became famous precisely by criticizing majority society. That's *sort of* like America being great, in that I think the parts of it that point out how bad the rest of it are often make excellent points. Vote for me!"

(sorry, I make fun of you because I love you)

There was a big brouhaha a couple of years ago when, as it first became apparent Obama had a good shot at the Presidency, Michelle Obama said that "for the first time in my adult life, I am proud of my country."

Republicans pounced on the comment, asking why she hadn't felt proud before, and she backtracked saying of course she was proud all the time and she loves America with the burning fury of a million suns and she was just saying that the Obama campaign was *particularly* inspiring.

As unconvincing denials go, this one was pretty far up there. But no one really held it against her. Probably most Obama voters felt vaguely the same way. *I* was an Obama voter, and I have proud memories of spending my Fourth of Julys as a kid debunking people's heartfelt emotions of patriotism. Aaron Sorkin:

> [What makes America the greatest country in the world?] It's not the greatest country in the world! We're seventh in literacy, 27th in math, 22nd in science,

> 49th in life expectancy, 178th in infant mortality, third in median household income, No. 4 in labor force, and No. 4 in exports. So when you ask what makes us the greatest country in the world, I don't know what the f*** you're talking about.

(Another good retort is "We're number one? Sure – number one in incarceration rates, drone strikes, and making new parents go back to work!")

All of this is true, of course. But it's weird that it's such a classic interest of members of the Blue Tribe, and members of the Red Tribe never seem to bring it up.

("We're number one? Sure – number one in levels of sexual degeneracy! Well, I guess probably number two, after the Netherlands, but they're really small and shouldn't count.")

My hunch – both the Red Tribe and the Blue Tribe, for whatever reason, identify "America" with the Red Tribe. Ask people for typically "American" things, and you end up with a very Red list of characteristics – guns, religion, barbecues, American football, NASCAR, cowboys, SUVs, unrestrained capitalism.

That means the Red Tribe feels intensely patriotic about "their" country, and the Blue Tribe feels like they're living in fortified enclaves deep in hostile territory.

Here is a popular piece published on a major media site called America: A Big, Fat, Stupid Nation. Another: America: A Bunch Of Spoiled, Whiny Brats. Americans are ignorant, scientifically illiterate religious fanatics whose "patriotism" is actually just narcissism. You Will Be Shocked At How Ignorant Americans Are, and we should Blame The Childish, Ignorant American People.

Needless to say, every single one of these articles was written by an American and read almost entirely by Americans. Those Americans very likely enjoyed the articles very much and did not feel the least bit insulted.

And look at the sources. HuffPo, Salon, Slate. Might those have anything in common?

On both sides, "American" can be either a normal demonym, or a code word for a member of the Red Tribe.

VIII.

The other day, I logged into OKCupid and found someone who looked cool. I was reading over her profile and found the following sentence:

> Don't message me if you're a sexist white guy

And my first thought was "Wait, so a sexist black person would be okay? Why?"

(The girl in question was white as snow)

Around the time the Ferguson riots were first starting, there were a host of articles with titles like Why White People Don't Seem To Understand Ferguson, Why It's So Hard For Whites To Understand Ferguson, and White Folks Listen Up And Let Me Tell You What Ferguson Is All About, this last of which says:

> Social media is full of people on both sides making presumptions, and believing what they want to believe. But it's the white folks that don't understand what this is all about. Let me put it as simply as I can for you […]

> No matter how wrong you think Trayvon Martin or Michael Brown were, I think we can all agree they didn't deserve to die over it. I want you white folks to understand that this is where the anger is coming from. You focused on the looting…."

And on a hunch I checked the author photos, and every single one of these articles was written by a white person.

White People Are Ruining America? White. White People Are Still A Disgrace? White. White Guys: We Suck And We're Sorry? White. Bye Bye, Whiny White Dudes? White. Dear Entitled Straight White Dudes, I'm Evicting You From My Life? White. White Dudes Need To Stop Whitesplaining? White. Reasons Why Americans Suck #1: White People? White.

We've all seen articles and comments and articles like this. Some unsavory people try to use them to prove that white people are the *real* victims or the media is biased against white people or something. Other people who are very nice and optimistic use them to show that some white people have developed some self-awareness and are willing to engage in self-criticism.

But I think the situation with "white" is much the same as the situation with "American" – it can either mean what it says, or be a code word for the Red Tribe.

(except on the blog Stuff White People Like, where it obviously serves as a code word for the *Blue* tribe. I don't know, guys. I didn't do it.)

I realize that's making a strong claim, but it would hardly be without precedent. When people say things like "gamers are misogynist", do they mean the 52% of gamers who are women? Do they mean every one of the 59% of Americans from every walk of life who are known to play video or computer games occasionally? No. "Gamer" is a coded reference to the Gray Tribe, the half-branched-off collection of libertarianish tech-savvy nerds, and everyone knows it. As well expect that when people talk about "fedoras", they mean Indiana Jones. Or when they talk about "urban youth", they mean freshmen at NYU. Everyone knows exactly who we mean when we say "urban youth", and them being young people who live in a city has only the most tenuous of relations to the actual concept.

And I'm saying words like "American" and "white" work the same way. Bill Clinton was the "first black President", but if Herman Cain had won in 2012 he'd have been the 43rd white president. And when an angry white person talks at great length about how much he hates "white dudes", *he is not being humble and self-critical*.

IX.

Imagine hearing that a liberal talk show host and comedian was so enraged by the actions of ISIS that he'd recorded and posted a video in which he shouts at them for ten minutes, cursing the "fanatical terrorists" and calling them "utter savages" with "savage values".

If *I* heard that, I'd be kind of surprised. It doesn't fit my model of what liberal talk show hosts do.

But the story I'm *actually* referring to is liberal talk show host / comedian Russell Brand making that same rant against Fox News for *supporting war against* the Islamic State, adding at the end that "Fox is worse than ISIS".

That fits my model perfectly. You wouldn't celebrate Osama's death, only Thatcher's. And you wouldn't call ISIS savages, only Fox News. Fox is the outgroup, ISIS is just some random people off in a desert. You hate the outgroup, you don't hate random desert people.

I would go further. Not only does Brand not feel much like hating ISIS, he has a strong incentive not to. That incentive is: the Red Tribe is known to hate ISIS loudly and conspicuously. Hating ISIS would signal Red Tribe membership, would be the equivalent of going into Crips territory with a big Bloods gang sign tattooed on your shoulder.

But this might be unfair. What would Russell Brand answer, if we asked him to justify his decision to be much angrier at Fox than ISIS?

He might say something like "Obviously Fox News is not literally worse than ISIS. But here I am, talking to my audience, who are mostly white British people and Americans. These people already know that ISIS is bad; they don't need to be told that any further. In fact, at this point being angry about how bad ISIS is, is less likely to genuinely change someone's mind about ISIS, and more likely to promote Islamophobia. The sort of people in my audience are at zero risk of becoming ISIS supporters, but at a very real risk of Islamophobia. So ranting against ISIS would be counterproductive and dangerous.

On the other hand, my audience of white British people and Americans is very likely to contain many Fox News viewers and supporters. And Fox, while not quite as evil as ISIS, is still pretty bad. So here's somewhere I have a genuine chance to reach people at risk and change minds. Therefore, I think my decision to rant against Fox News, and maybe hyperbolically say they were 'worse than ISIS' is justified under the circumstances."

I have a lot of sympathy to hypothetical-Brand, especially to the part about Islamophobia. It *does* seem really possible to denounce ISIS' atrocities to a population that already hates them in order to weak-man a couple of already-marginalized Muslims. We need to fight terrorism and atrocities – therefore it's okay to shout at a poor girl ten thousand miles from home for wearing a headscarf in public. Christians are being executed for their faith in Sudan, therefore let's picket the people trying to build a mosque next door.

But my sympathy with Brand ends when he acts like his audience is likely to be fans of Fox News.

In a world where a negligible number of Redditors oppose gay marriage and 1% of Less Wrongers identify conservative and I know 0/150 creationists, how many of the people who visit the YouTube channel of a well-known liberal activist with a Che-inspired banner, a channel whose episode names are things like "War: What Is It Good For?" and "Sarah Silverman Talks Feminism" – how many of them do you think are big Fox News fans?

In a way, Russell Brand would have been *braver* taking a stand against ISIS than against Fox. If he attacked ISIS, his viewers would just be a little confused and uncomfortable. Whereas every moment he's attacking Fox his viewers are like "HA HA! YEAH! GET 'EM! SHOW THOSE IGNORANT BIGOTS IN THE OUTGROUP WHO'S BOSS!"

Brand acts as if there are just these countries called “Britain” and “America” who are receiving his material. Wrong. There are two parallel universes, and he’s only broadcasting to one of them.

The result is exactly what we predicted would happen in the case of Islam. Bombard people with images of a far-off land they already hate and tell them to hate it more, and the result is ramping up the intolerance on the couple of dazed and marginalized representatives of that culture who have ended up stuck on your half of the divide. Sure enough, if industry or culture or community gets Blue enough, Red Tribe members start getting harassed, fired from their jobs (Brendan Eich being the obvious example) or otherwise shown the door.

Think of Brendan Eich as a member of a tiny religious minority surrounded by people who hate that minority. Suddenly firing him doesn’t seem very noble.

If you mix together Podunk, Texas and Mosul, Iraq, you can prove that Muslims are scary and very powerful people who are executing Christians all the time – and so we have a great excuse for kicking the one remaining Muslim family, random people who never hurt anyone, out of town.

And if you mix together the open-source tech industry and the parallel universe where you can’t wear a FreeBSD t-shirt without risking someone trying to exorcise you, you can prove that Christians are scary and very powerful people who are persecuting everyone else all the time, and you have a great excuse for kicking one of the few people willing to affiliate with the Red Tribe, a guy who never hurt anyone, out of town.

When a friend of mine heard Eich got fired, she didn’t see anything wrong with it. “I can tolerate anything except intolerance,” she said.

“Intolerance” is starting to look like another one of those words like “white” and “American”.

“I can tolerate anything except the outgroup.” Doesn’t sound quite so noble now, does it?

X.

We started by asking: millions of people are conspicuously praising every outgroup they can think of, while conspicuously condemning their own in-group. This seems contrary to what we know about social psychology. What’s up?

We noted that outgroups are rarely literally “the group most different from you”, and in fact far more likely to be groups very similar to you sharing *almost* all your characteristics and living in the same area.

We then noted that although liberals and conservatives live in the same area, they might as well be two totally different countries or universe as far as level of interaction were concerned.

Contra the usual idea of them being marked only by voting behavior, we described them as very different tribes with totally different cultures. You can speak of “American culture” only in the same way you can speak of “Asian culture” – that is, with a lot of interior boundaries being pushed under the rug.

The outgroup of the Red Tribe is occasionally blacks and gays and Muslims, more often the Blue Tribe.

The Blue Tribe has performed some kind of very impressive act of alchemy, and transmuted *all* of its outgroup hatred to the Red Tribe.

This is not surprising. Ethnic differences have proven quite tractable in the face of shared strategic aims. Even the Nazis, not known for their ethnic tolerance, were able to get all buddy-buddy with the Japanese when they had a common cause.

Research suggests Blue Tribe / Red Tribe prejudice to be much stronger than better-known types of prejudice like racism. Once the Blue Tribe was able to enlist the blacks and gays and Muslims in their ranks, they became allies of convenience who deserve to be rehabilitated with mildly condescending paeans to their virtue. “There never was a coward where the shamrock grows.”

Spending your entire life insulting the other tribe and talking about how terrible they are makes you look, well, tribalistic. It is definitely not high class. So when members of the Blue Tribe decide to dedicate their entire life to yelling about how terrible the Red Tribe is, they make sure that instead of saying “the Red Tribe”, they say “America”, or “white people”, or “straight white men”. That way it’s *humble self-criticism*. They are *so* interested in justice that they are willing to critique *their own beloved side*, much as it pains them to do so. We know they are not exaggerating, because one might exaggerate the flaws of an enemy, but that anyone would exaggerate their *own* flaws fails the criterion of embarrassment.

The Blue Tribe always has an excuse at hand to persecute and crush any Red Tribers unfortunate enough to fall into its light-matter-universe by defining them as all-powerful domineering oppressors. They appeal to the fact that this is definitely the way it works in the Red Tribe’s dark-matter-universe, and that’s in the same country so it has to be the same community for all intents and purposes. As a result, every Blue Tribe institution is permanently licensed to take whatever emergency measures are necessary against the Red Tribe, however disturbing they might otherwise seem.

And so how virtuous, how noble the Blue Tribe! Perfectly tolerant of all of the different groups that just so happen to be allied with them, never intolerant unless it happen to be against intolerance itself. Never stooping to engage in petty tribal conflict like that awful Red Tribe, but always nobly criticizing their own culture and striving to make it better!

Sorry. But I hope this is at least a *little* convincing. The weird dynamic of outgroup-philia and ingroup-phobia isn’t anything of the sort. It’s just good old-fashioned in-group-favoritism and outgroup bashing, a little more sophisticated and a little more sneaky.

XI.

This essay is bad and I should feel bad.

I should feel bad because I made *exactly* the mistake I am trying to warn everyone else about, and it wasn’t until I was almost done that I noticed.

How virtuous, how noble I must be! Never stooping to engage in petty tribal conflict like that silly Red Tribe, but always nobly criticizing my own tribe and striving to make it better.

Yeah. Once I've written a ten thousand word essay savagely attacking the Blue Tribe, either I'm a very special person or they're my outgroup. And I'm not *that* special.

Just as you can pull a fast one and look humbly self-critical if you make your audience assume there's just one American culture, so maybe you can trick people by assuming there's only one Blue Tribe.

I'm pretty sure I'm not Red, but I did talk about the Grey Tribe above, and I show all the risk factors for being one of them. That means that, although my critique of the Blue Tribe may be right or wrong, in terms of *motivation* it comes from the same place as a Red Tribe member talking about how much they hate al-Qaeda or a Blue Tribe member talking about how much they hate ignorant bigots. And when I boast of being able to tolerate Christians and Southerners whom the Blue Tribe is mean to, I'm not being tolerant at all, just noticing people so far away from me they wouldn't make a good outgroup anyway.

I had *fun* writing this article. People do not have fun writing articles savagely criticizing their in-group. People can criticize their in-group, it's not *humanly impossible*, but it takes nerves of steel, it makes your blood boil, you should sweat blood. It shouldn't be *fun*.

You can bet some white guy on Gawker who week after week churns out "Why White People Are So Terrible" and "Here's What Dumb White People Don't Understand" is having fun and not sweating any blood at all. He's not criticizing his in-group, he's never even *considered* criticizing his in-group. I can't blame him. Criticizing the in-group is a really difficult project I've barely begun to build the mental skills necessary to even consider.

I can think of criticisms of my own tribe. Important criticisms, true ones. But the thought of writing them makes my blood boil.

I imagine might I feel like some liberal US Muslim leader, when he goes on the O'Reilly Show, and O'Reilly ambushes him and demands to know why he and other American Muslims haven't condemned beheadings by ISIS more, demands that he criticize them right there on live TV. And you can see the wheels in the Muslim leader's head turning, thinking something like "Okay, obviously beheadings are terrible and I hate them as much as anyone. But you don't care even *the slightest bit* about the victims of beheadings. You're just looking for a way to score points against me so you can embarass all Muslims. And I would rather personally behead every single person in the world than give a smug bigot like you a single microgram more stupid self-satisfaction than you've already got."

That is how I feel when asked to criticize my own tribe, even for correct reasons. If you think you're criticizing your own tribe, and your blood is not at that temperature, consider the possibility that you aren't.

But if I want Self-Criticism Virtue Points, criticizing the Grey Tribe is the only honest way to get them. And if I want Tolerance Points, my own personal cross to bear right now is tolerating the Blue Tribe. I need to remind myself that when they are bad people, they are merely Osama-level bad people instead of Thatcher-level bad people. And when they are good people, they are powerful and necessary crusaders against the evils of the world.

The worst thing that could happen to this post is to have it be used as convenient feces to fling at the Blue Tribe whenever feces are necessary. Which, given what has

happened to my last couple of posts along these lines and the obvious biases of my own subconscious, I already expect it will be.

But the best thing that could happen to this post is that it makes a lot of people, especially myself, figure out how to be more tolerant. Not in the “of course I’m tolerant, why shouldn’t I be?” sense of the Emperor in Part I. But in the sense of “being tolerant makes me see red, makes me sweat blood, but darn it *I am going to be tolerant anyway*.”

Happy Petrov Day! If Today is Also Your Birthday, Happy Birthday!

This is a linkpost for https://akataskeuastos.wordpress.com/2017/09/26/happy-petrov-day-if-today-is-also-your-birthday-happy-birthday/

A discussion of September 26 and September 1993, and an invitation to join the club if that's your birthday too

Nobody does the thing that they are supposedly doing

I feel like one of the most important lessons I've had about How the World Works, which has taken quite a bit of time to sink in, is:

In general, neither organizations nor individual people do the thing that their supposed role says they should do. Rather they tend to do the things that align with their incentives (which may sometimes be economic, but even more often they are social and psychological). If you want to really change things, you have to change people's incentives.

But I feel like I've had to gradually piece this together from a variety of places, over a long time; I've never read anything that would have laid down the whole picture. I remember that Freakonomics had a few chapters about how incentives cause unexpected behavior, but that was mostly about economic incentives, which are just a small part of the whole picture. And it didn't really focus on the "nothing in the world works the way you'd naively expect" thing; as I recall, it was presented more as a curiosity.

On the other hand, Robin Hanson has had a lot of stuff about "X is not about Y", but that has mostly been framed in terms of prestige and signaling, which is the kind of stuff that's certainly an important part of the whole picture (the psychological kind of incentives), but again just a part of the picture. (However, his upcoming book goes into a lot more detail on why and how the publicly-stated motives for human or organizational behavior aren't actually the true motives.)

And then in social/evolutionary/moral psychology there's a bunch of stuff about social-psychological incentives, of how we're motivated to denounce outgroups and form bonds with our ingroups; and how it can be socially costly to have accurate beliefs about outgroups and defend them to your ingroup, whereas it would be much more rewarding to just spread inaccuracies or outright lies about how terrible the outgroups are, and thus increase your own social standing. And how even well-meaning ideologies will by default get hijacked by these kinds of dynamics and become something quite different from what they claimed to be.

But again, that's just one piece of the whole story. And you can find more isolated pieces of the whole story scattered around in a variety of articles and books, also stuff like the iron law of oligarchy, rational irrationality, public choice theory, etc etc. But no grand synthesis.

There's also a relevant strand of this in the psychology of motivation/procrastination/habit-formation, on why people keep putting off various things that they claim they *want* to do, but then don't. And how small things can reshape people's behavior, like if somebody ends up as a much more healthy eater just because they *don't* happen to have a fast food restaurant conveniently near their route home from work. Which isn't necessarily so much about incentives themselves, but an important building block in understanding why our behavior tends to be so strongly shaped by things that are entirely separate from consciously-set goals.

Additionally, the things that do drive human behavior are often things like maintaining a self-concept, seeking feelings of connection, autonomy and competence,

maintaining status, enforcing various moral intuitions, etc., things that only loosely align one's behavior with one's stated goals. Often people may not even realize what exactly it is that they are trying to achieve with their behavior.

"Experiental pica" is a misdirected craving for something that doesn't actually fulfill the need behind the craving. The term originally comes from a condition where people with a mineral deficiency start eating things like ice, which don't actually help with the deficiency. Recently I've been shifting towards the perspective that, to a first approximation, roughly *everything* that people do is pica for some deeper desire, with that deeper desire being something like social connection, feeling safe and accepted, or having a feeling of autonomy or competence. That is, most of the things that people will give as reasons for why they are doing something will actually miss the mark, and also that many people are engaging in things that are actually relatively inefficient ways of achieving their true desires, such as pursuing career success when the real goal is social connection. (This doesn't mean that the underlying desire would never be fulfilled, just that it gets fulfilled less often than it would if people were aware of their true desires.)

In Defense of Unreliability

In a long post mostly about a different issue, Zvi Mowshowitz writes:

> I also strongly endorse that the default level of reliability needs to be much, much higher than the standard default level of reliability, especially in The Bay. Things there are really bad.
>
> When I make a plan with a friend in The Bay, I never assume the plan will actually happen. There is actual no one there I feel I can count on to be on time and not flake. I would come to visit more often if plans could actually be made. Instead, suggestions can be made, and half the time things go more or less the way you planned them. This is a terrible, very bad, no good equilibrium. Are there people I want to see badly enough to put up with a 50% reliability rate? Yes, but there are not many, and I get much less than half the utility out of those friendships than I would otherwise get.

First of all, I'd like to say that nothing in my post should be construed as saying Zvi's desire for reliable friends is invalid or wrong. It's disappointing to expect a friend to come over and then they don't. If you're a busy person, on vacation or otherwise limited in time, a friend's canceled plans may mean that you've missed out on an important opportunity to do something productive and/or fun. It is very reasonable to want to befriend people who will reliably show up places they said they will on time. However, I do want to explain why I myself am quite unreliable and how I benefit from a social norm in which this unreliability is acceptable. (We should also note that I have lived in the Bay for the majority of my adult, actually-socializing life, so I may be unfamiliar with the benefits of a non-flake lifestyle.)

I primarily get places through public transit and Uberpool. The Bay Area's public transit system is really really good compared to public transit in most of the rest of the country (for one thing, it is possible to get places on it). However, our public transit is certainly inferior to, say, New York City's. One of the ways this works is that sometimes, based on the Inscrutable Whim of the Train Gods, the train will choose to show up fourteen minutes late. Uberpool also has high variance in time estimates, because they have to pick up and drop off other people. What this means is that when I say "I will get there at such-and-such a time", I mean "there is a bimodal distribution of times when I could show up which is centered around this time and probably has a standard deviation of like five to ten minutes."

So there are ways I can fairly consistently show up on time. One is that I could take UberX wherever I'm going and eat the extra expense-- although doing that consistently would trade off against my goal of using money responsibly. Another is that I can plan to show up on average ten or fifteen minutes before I'm supposed to show up, and then most of the time I will be on time. (This is what I do for doctors' and therapists' appointments.)

There are two problems with adopting the latter strategy in general. First, my time also has value! If it's bad for me to show up ten minutes late because the person is waiting around being bored, then it is also bad for me to show up ten minutes early so I have to wait around and be bored. Second, in many cases, showing up early is just as inconvenient for others as showing up late. For instance, if a friend invited me over for dinner and I show up fifteen minutes early, they might be still in their bathrobe and

really counting on that fifteen minutes to shove the floordrobe into the closet and take the garbage out. That would be considerably ruder than showing up fifteen minutes late (at least if you keep them posted), because at that point the food is probably only beginning to get cold.

(I guess I could arrive early and then hang out on a street corner until it was time for dinner but see above re: my time has value.)

In general, instead of trying to always show up before you said you would, I think the best strategy is to try to be early about as often as you are late, unless it is something where being early is much much better than being late (a theatrical production, a doctor's appointment, a job interview) or vice versa (a party with lots of other invitees).

However, Zvi didn't just talk about being on time: he also talked about flaking. My local corner of the Bay seems to have less of a flaking problem than his corner. I, a diagnosed agoraphobe, still manage to make the majority of the social events I agree to go to, and many people of my acquaintance make as much as ninety or ninety-five percent. (Maybe I am particularly charming and people don't want to flake on me, or maybe I'm proactive and flake on them first.) But I think it is very useful that no one gets angry at me for flaking as much as I do.

I'm scared of leaving my house. This means that when I make social arrangements a lot of the time I won't end up actually going to them because I will be too scared of leaving my house. Whether I'm going to have a good mental health day or a bad mental health day is hard to predict even a week in advance, because it depends on short-term triggers like whether I've fought with a close friend, whether the assholes across the street have decided to set off fireworks, whether a person has said something unpleasant about me on the Internet, whether I've been doing a good job of remembering that in spite of what my brain tells me doing things will make me feel better and not doing things will make me feel worse, and so on. So the only way I can achieve any sort of reliability in social arrangements is by not making them.

I do not want to not make social arrangements. Social isolation makes my mental health worse. And doing literally anything tends to make me less depressed. I am also informed that some people would occasionally like to talk to me [citation needed]. So therefore I have decided to make plans anyway, and push onto my friends the negative consequences of dealing with my flakiness.

It seems perfectly reasonable to me that one would object to this state of affairs and choose not to have me as a friend. (This is one of many good reasons why someone might not want to have me as a friend.) But I think before advocating for a complete shift in social norms one should consider the benefits the social norms already have to those participating in them.

Intellectual Progress Inside and Outside Academia

This post is taken from a recent facebook conversation that included Wei Dai, Eliezer Yudkowsky, Vladimir Slepnev, Stuart Armstrong, Maxim Kesin, Qiaochu Yuan and Robby Bensinger, about the ability of academia to do the key intellectual progress required in AI alignment.

[The above people all gave permission to have their comments copied here. Some commenters requested their replies not be made public, and their comment threads were not copied over.]

Initial Thread

Wei Dai:

> Eliezer, can you give us your take on this discussion between me, Vladimir Slepnev, and Stuart Armstrong? I'm especially interested to know if you have any thoughts on what is preventing academia from taking or even recognizing certain steps in intellectual progress (e.g., inventing anything resembling Bitcoin or TDT/UDT) that non-academics are capable of. What is going on there and what do we need do to avoid possibly suffering the same fate? See this and this.

Eliezer Yudkowsky:

> It's a deep issue. But stating the obvious is often a good idea, so to state the obvious parts, we're looking at a lot of principal-agent problems, Goodhart's Law, bad systemic incentives, hypercompetition crowding out voluntary contributions of real work, the blind leading the blind and second-generation illiteracy, etcetera. There just isn't very much in the academic system that does promote any kind of real work getting done, and a lot of other rewards and incentives instead. If you wanted to get productive work done inside academia, you'd have to ignore all the incentives pointing elsewhere, and then you'd (a) be leading a horrible unrewarded life and (b) you would fall off the hypercompetitive frontier of the major journals and (c) nobody else would be particularly incentivized to pay attention to you except under unusual circumstances. Academia isn't about knowledge. To put it another way, although there are deep things to say about the way in which bad incentives arise, the skills that are lost, the particular fallacies that arise, and so on, it doesn't feel to me like the *obvious* bad incentives are inadequate to explain the observations you're pointing to. Unless there's some kind of psychological block preventing people from seeing all the obvious systemic problems, it doesn't feel like the end result ought to be surprising.
>
> Of course, a lot of people do seem to have trouble seeing what I'd consider to be obvious systemic problems. I'd chalk that up to not as much fluency with Moloch's toolbox, plus them not being status-blind and assigning non-zero positive status to academia that makes them emotionally reluctant to correctly take all the obvious problems at face value.

Eliezer Yudkowsky (cont.):

> It seems to me that I've watched organizations like OpenPhil try to sponsor academics to work on AI alignment, and it seems to me that they just can't produce what I'd consider to be real work. The journal paper that Stuart Armstrong coauthored on "interruptibility" is a far step down from Armstrong's other work on corrigibility. It had to be dumbed way down (I'm counting obscuration with fancy equations and math results as "dumbing down") to be published in a mainstream journal. It had to be stripped of all the caveats and any mention of explicit incompleteness, which is necessary meta-information for any ongoing incremental progress, not to mention important from a safety standpoint. The root cause can be debated but the observable seems plain. If you want to get real work done, the obvious strategy would be to not subject yourself to any academic incentives or bureaucratic processes. Particularly including peer review by non-"hobbyists" (peer commentary by fellow "hobbyists" still being potentially very valuable), or review by grant committees staffed by the sort of people who are still impressed by academic sage-costuming and will want you to compete against pointlessly obscured but terribly serious-looking equations.

Eliezer Yudkowsky (cont.):

> There's a lot of detailed stories about good practice and bad practice, like why mailing lists work better than journals because of that thing I wrote on FB somewhere about why you absolutely need 4 layers of conversation in order to have real progress and journals do 3 layers which doesn't work. If you're asking about those it's a lot of little long stories that add up.

Subthread 1

Wei Dai:

> Academia is capable of many deep and important results though, like complexity theory, public-key cryptography, zero knowledge proofs, vNM and Savage's decision theories, to name some that I'm familiar with. It seems like we need a theory that explains why it's able to take certain kinds of steps but not others, or maybe why the situation has gotten a lot worse in recent decades.
>
> That academia may not be able to make progress on AI alignment is something that worries me and a major reason for me to be concerned about this issue now. If we had a better, more nuanced theory of what is wrong with academia, that would be useful for guiding our own expectations on this question and perhaps also help persuade people in charge of organizations like OpenPhil.

Qiaochu Yuan:

> Public-key cryptography was invented by GCHQ first, right?

Wei Dai:

> It was independently reinvented by academia, with only a short delay (4 years according to Wikipedia) using much less resources compared to the government agencies. That seems good enough to illustrate my point that academia is (or at least was) capable of doing good and efficient work.

Qiaochu Yuan:

> Fair point.
>
> I'm a little concerned about the use of the phrase "academia" in this conversation not cutting reality at the joints. Academia may simply not be very homogeneous over space and time - it certainly seems strange to me to lump von Neumann in with everyone else, for example.

Wei Dai:

> Sure, part of my question here is how to better carve reality at the joints. What's the relevant difference between the parts (in space and/or time) of academia that are productive and the parts that are not?

Stuart Armstrong:

> Academia is often productive. I think the challenge is mainly getting it to be productive on the right problems.

Wei Dai:

> Interesting, so maybe a better way to frame my question is, of the times that academia managed to focus on the right problems, what was responsible for that? Or, what is causing academia to not be able to focus on the right problems in certain fields now?

Subthread 2

Eliezer Yudkowsky:

> Things have certainly gotten a lot worse in recent decades. There's various stories I've theorized about that but the primary fact seems pretty blatant. Things might be different if we had the researchers and incentives from the 1940s, but modern academics are only slightly less likely to sprout wings than to solve real alignment problems as opposed to fake ones. They're still the same people and the same incentive structure that ignored the entire issue in the first place.
>
> OpenPhil is better than most funding sources, but not close to adequate. I model them as having not seen past the pretend. I'm not sure that more nuanced theories are what they need to break free. Sure, I have a dozen theories about various factors. But ultimately, most human institutions through history haven't solved hard mental problems. Asking why modern academia doesn't UDT may be like asking why JC Penney doesn't. It's just not set up to do that. Nobody is being docked a bonus for writing papers about CDT instead. Feeling worried and like something is out of place about the College of Cardinals in the Catholic Church not inventing cryptocurrencies, suggests a basic mental tension that may not be cured by more nuanced theories of the sociology of religion. Success is unusual and calls for explanation, failure doesn't. Academia in a few colleges in a few countries used to be in a weird regime where it could solve hard problems, times changed, it fell out of that weird place.

Rob Bensinger:

It's not actually clear to me, even after all this discussion, that 1940s researchers had significantly better core mental habits / mindsets for alignment work than 2010s researchers. A few counter-points:

- A lot of the best minds worked on QM in the early 20th century, but I don't see clear evidence that QM progressed differently than AI is progressing today; that is, I don't know of a clear case that falsifies the hypothesis "all the differences in output are due to AI and QM as cognitive problems happening to involve inherently different kinds and degrees of difficulty". In both cases, it seems like people did a good job of applying conventional scientific methods and occasionally achieving conceptual breakthroughs in conventional scientific ways; and in both cases, it seems like there's a huge amount of missing-the-forest-for-the-trees, not-seriously-thinking-about-the-implications-of-beliefs, and generally-approaching-philosophyish-questions-flippantly. It took something like 50 years to go from "Schrodinger's cat is weird" to "OK /maybe/ macroscopic superposition-ish things are real" in physics, and "maybe macroscopic superposition-ish things are real" strikes me as much more obvious and much less demanding of sustained theorizing than, e.g., 'we need to prioritize decision theory research ASAP in order to prevent superintelligent AI systems from destroying the world'. Even von Neumann had non-naturalist views about QM, and if von Neumann is a symptom of intellectual degeneracy then I don't know what isn't.

- Ditto for the development of nuclear weapons. I don't see any clear examples of qualitatively better forecasting, strategy, outside-the-box thinking, or scientific productivity on this topic in e.g. the 1930s, compared to what I'd expect see today. (Though this comparison is harder to make because we've accumulated a lot of knowledge and hard experience with technological GCR as a result of this and similar cases.) The near-success of the secrecy effort might be an exception, since that took some loner agency and coordination that seems harder to imagine today. (Though that might also have been made easier by the smaller and less internationalized scientific community of the day, and by the fact that world war was on everyone's radar?)

- Turing and I. J. Good both had enough puzzle pieces to do at least a little serious thinking about alignment, and there was no particular reason for them not to do so. The 1956 Dartmouth workshop shows "maybe true AI isn't that far off" was at least taken somewhat seriously by a fair number of people (though historians tend to overstate the extent to which this was true). If 1940s researchers were dramatically better than 2010s researchers at this kind of thing, and the decay after the 1940s wasn't instantaneous, I'd have expected at least a hint of serious thinking-for-more-than-two-hours about alignment from at least one person working in the 1950s-1960s (if not earlier).

Rob Bensinger:

Here's a different hypothesis: Human brains and/or all of the 20th century's standard scientific toolboxes and norms are just really bad at philosophical/conceptual issues, full stop. We're bad at it now, and we were roughly equally bad at it in the 1940s. A lot of fields have slowed down because we've plucked most of the low-hanging fruit that doesn't require deep philosophical/conceptual innovation, and AI in particular happens to be an area

> where the things human scientists have always been worst at are especially critical for success.

Wei Dai:

> Ok, so the story I'm forming in my mind is that we've always been really bad at philosophical/conceptual issues, and past philosophical/conceptual advances just represent very low-hanging fruit that have been picked. When we invented mailing lists / blogs, the advantage over traditional academic communications allowed us to reach a little higher and pick up a few more fruits but progress is still very limited because we're still not able to reach very high in an absolute sense, and making progress this way depends on gathering together enough hobbyists with the right interests and resources which is a rare occurrence. Rob, I'm not sure how much of this you endorse, but it seems like the best explanation of all the relevant facts I've seen so far.

Rob Bensinger:

> I think the object-level philosophical progress via mailing lists / blogs was tied to coming up with some good philosophical methodology. One simple narrative about the global situation (pretty close to the standard narrative) is that before 1880 or so, human inquiry was good at exploring weird nonstandard hypotheses, but bad at rigorously demanding testability and precision of those hypotheses. Human inquiry between roughly 1880 and 1980 solved that problem by demanding testability and precision in all things, which (combined with prosaic knowledge accumulation) let them grab a lot of low-hanging scientific fruit really fast, but caused them to be unnecessarily slow at exploring any new perspectives that weren't 100% obviously testable and precise in a certain naive sense (which led to lack-of-serious-inquiry into "weird" questions at the edges of conventional scientific activities, like MWI and Newcomb's problem).
>
> Bayesianism, the cognitive revolution, the slow fade of positivism's influence, the random walk of academic one-upmanship, etc. eventually led to more sophistication in various quarters about what kind of testability and precision are important by the late 20th century, but this process of synthesizing 'explore weird nonstandard hypotheses' with 'demand testability and precision' (which are the two critical pieces of the puzzle for 'do unusually well at philosophy/forecasting/etc.') was very uneven and slow. Thus you get various little islands of especially good philosophy-ish thinking showing up at roughly the same time here and there, including parts of analytic philosophy (e.g., Drescher), mailing lists (e.g., Extropians), and psychology (e.g., Tetlock).

Subthread 3

Vladimir Slepnev:

> Eliezer, your position is very sharp. A couple questions then:
>
> 1. Do you think e.g. Scott Aaronson's work on quantum computing today falls outside the "weird regime where it could solve hard problems"?
>
> 2. Do you have a clear understanding why e.g. Nick Bostrom isn't excited about TDT/UDT?

Wei Dai:

> Vladimir, can you clarify what you mean by "isn't excited"? Nick did write a few paragraphs about the relevance of decision theory to AI alignment in his Superintelligence, and cited TDT and UDT as "newer candidates [...] which are still under development". I'm not sure what else you'd expect, given that he hasn't specialized in decision theory in his philosophy work? Also, what's your own view of what's causing academia to not be able to make these "outsider steps"?

Vladimir Slepnev:

> Wei, at some point you thought of UDT as the solution to anthropic reasoning, right? That's Bostrom's specialty. So if you are right, I'd expect more than a single superficial mention.
>
> My view is that academia certainly tends to go off in wrong directions and it was always like that. But its direction can be influenced with enough effort and understanding, it's been done many times, and the benefits of doing that are too great to overlook.

Wei Dai:

> I'm not sure, maybe he hasn't looked into UDT closely enough to understand the relevance to anthropics or he's committed to a probability view? Probably Stuart has a better idea of this than I do. Oh, I do recall that when I attended a workshop at FHI, he asked me some questions about UDT that seemed to indicate that he didn't understand it very well. I'm guessing he's probably just too busy to do object-level philosophical investigations these days.
>
> Can you give some past examples of academia going off in the wrong direction, and that being fixed by outsiders influencing its direction?

Vladimir Slepnev:

> Why do you need the "fixed by outsiders" bit? I think it's easier to change the direction of academia while being in academia, and that's been done many times.

Maxim Kesin:

> Vladimir Slepnev The price of admission is pretty high for people who can do otherwise productive work, no? Especially since very few members of the club can have direction-changing impact. Something like finding and convincing existing high-standing members, preferably several of them seems like a better strategy than joining the club and doing it from the inside yourself.

Wei Dai:

> Vladimir, on LW you wrote "More like a subset of steps in each field that need to be done by outsiders, while both preceding and following steps can be done by academia." If some academic field is going in a wrong direction because it's missing a step that needs to be done by outsiders, how can someone in academia change its direction? I'm confused... Are you saying outsiders should go into academia in order to change its direction, after taking the missing "outsider steps"? Or that there is no direct past evidence that outsiders can change

academia's direction but there's evidence that insiders can and that serves as bayesian evidence that outsiders can too? Or something else?

Vladimir Slepnev:

I guess I shouldn't have called them "outsider steps", more like "newcomer steps". Does that make sense?

Eliezer Yudkowsky:

There's an old question, "What does the Bible God need to do for the Christians to say he is not good?" What would academia need to do before you let it go?

Vladimir Slepnev:

But I don't feel abused! My interactions with academia have been quite pleasant, and reading papers usually gives me nice surprises. When I read your negative comments about academia, I mostly just get confused. At least from what I've read in this discussion today, it seems like the mystical force that's stopping people like Bostrom from going fully on board with ideas like UDT is simple miscommunication on our part, not anything more sinister. If our arguments for using decisions over probabilities aren't convincing enough, perhaps we should work on them some more.

Wei Dai:

Vladimir, surely those academic fields have had plenty of infusion of newcomers in the form of new Ph.D. students, but the missing steps only got done when people tried do them while remaining entirely out of academia. Are you sure the relevant factor here is "new to the field" rather than "doing work outside of academia"?

Stuart Armstrong:

Academic fields are often productive, but narrow. Saying "we should use decision theory instead of probability to deal with anthropics" falls outside of most of the relevant fields, so few academics are interested, because it doesn't solve the problems they are working on.

Wei Dai:

Vladimir, a lot of people on LW didn't have much trouble understanding UDT as informally presented there, or recognizing it as a step in the right direction. If joining academia makes somebody much less able to recognize progress in decision theory, that seems like a bad thing and we shouldn't be encouraging people to do that (at least until we figure out what exactly is causing the problem and how to fix or avoid it on an institutional or individual level).

Vladimir Slepnev:

I think it's not surprising that many LWers agreed with UDT, because most of them were introduced to the topic by Eliezer's post on Newcomb's problem, which framed the problem in a way that emphasized decisions over probabilities. (Eliezer, if you're listening, that post of yours was the single best example of persuasion I've seen in my life, and for a good goal too. Cheers!) So there's probably no statistical effect saying outsiders are better at grasping UDT on

> average. It's not that academia is lacking some decision theory skill, they just haven't bought our framing yet. When/if they do, they will be uniquely good at digging into this idea, just as with many other ideas.
>
> If the above is true, then refusing to pay the fixed cost of getting our ideas into academia seems clearly wrong. What do you think?

Subthread 4

Stuart Armstrong:

> Think the problem is a mix of specialisation and lack of urgency. If I'd been willing to adapt to the format, I'm sure I could have got my old pro-SIA arguments published. But anthropics wasn't ready for a "ignore the big probability debates you've been having; anthropic probability doesn't exist" paper. And those were interested in the fundamental interplay between probability and decision theory weren't interested in anthropics (and I wasn't willing to put the effort in to translate it into their language).
>
> This is where the lack of urgency comes in. People found the paper interesting, I'd wager, but not saying anything about the questions they were interested in. And they had no real feeling that some questions were far more important than theirs.

Stuart Armstrong:

> I've presented the idea to Nick a few times, but he never seemed to get it fully. It's hard to ignore probabilities when you've spent your life with them.

Eliezer Yudkowsky:

> I will mention for whatever it's worth that I don't think decision theory can eliminate anthropics. That's an intuition I still find credible and it's possible Bostrom felt the same. I've also seen Bostrom contribute at least one decision theory idea to anthropic problems, during a conversation with him by instant messenger, a division-of-responsibility principle that UDT later rendered redundant.

Stuart Armstrong:

> I also disagree with Eliezer about the use of the "interruptible agents" paper. The math is fun but ultimately pointless, and there is little mention of AI safety. However, it was immensely useful for me to write that paper with Laurent, as it taught me so much about how to model things, and how to try and translate those models into things that ML people like. As a consequence, I can now design indifference methods for practically any agent, which was not the case before.
>
> And of course the paper wouldn't mention the hard AI safety problems - not enough people in ML are working on those. The aim was to 1) present part of the problem, 2) present part of the solution, and 3) get both of those sufficiently accepted that harder versions of the problem can then be phrased as "take known problem/solution X, and add an extra assumption..."

Rob Bensinger:

> That rationale makes sense to me. I think the concern is: if the most visible and widely discussed papers in AI alignment continue to be ones that deliberately obscure their own significance in various ways, then the benefits from the slow build-up to being able to clearly articulate our actual views in mainstream outlets may be outweighed by the costs from many other researchers internalizing the wrong take-aways in the intervening time. This is particularly true if many different build-ups like this are occurring simultaneously, over many years of incremental progress toward just coming out and saying what we actually think.
>
> I think this is a hard problem, and one MIRI's repeatedly had to deal with. Very few of MIRI's academic publications even come close to giving a full rationale for why we care about a given topic or result. The concern is with making it standard practice for high-visibility AI alignment papers to be at least somewhat misleading (in order to get wider attention, meet less resistance, get published, etc.), rather than with the interruptibility paper as an isolated case; and this seems like a larger problem for overstatements of significance than for understatements.
>
> I don't know how best to address this problem. Two approaches MIRI has tried before, which might help FHI navigate this, are: (1) writing a short version of the paper for publication that doesn't fully explain the AI safety rationale, and a longer eprint of the same paper that does explain the rationale; and/or (2) explaining results' significance more clearly and candidly in the blog post announcing the paper.

Subthread 5

Eliezer Yudkowsky:

> To put this yet another way, most human bureaucracies and big organizations don't do science. They have incentives for the people inside them which get them to do things other than science. For example, in the FBI, instead of doing science, you can best advance your career by closing big-name murder cases... or whatever. In the field of psychology, instead of doing science, you can get a lot of undergraduates into a room and submit obscured-math impressive-sounding papers with a bunch of tables that claim a p-value greater than 0.05. Among the ways we know that this has little to do with science is that the papers don't replicate. P-values are rituals[1], and being surprised that the rituals don't go hand-in-hand with science says you need to adjust your intuitions about what is surprising. It's like being surprised that your prayers aren't curing cancer and asking how you need to pray differently.
>
> Now, it may be that separately from the standard incentives, decades later, a few heroes get together and try to replicate some of the most prestigious papers. They are doing science. Maybe somebody inside the FBI is also doing science. Lots of people in Christian religious organizations, over the last few centuries, did some science, though fewer now than before. Maybe the public even lauded the science they did, and they got some rewards. It doesn't mean the Catholic Church is set up to teach people how to do real science, or that this is the primary way to get ahead in the Catholic Church such that status-seekers will be driven to seek their promotions by doing great science.
>
> The people doing real science by trying to replicate psychology studies may report ritual p-values and submit for ritual peer-review-by-idiots. Similarly, some doctors

in the past no doubt prayed while giving their patients antibiotics. It doesn't mean that prayer works some of the time. It means that these heroes are doing science, and separately, doing bureaucracy and a kind of elaborate ritual that is what our generation considers to be prestigious and mysterious witch doctery.

[1] https://arbital.com/p/likelihoods_not_pvalues/?l=4x

Cognitive Empathy and Emotional Labor

This is a linkpost for https://mapandterritory.org/cognitive-empathy-and-emotional-labor-cb256c38597d

The concept of emotional labor has been popularized in the last couple years as a way of talking about the work people do to manage other people's emotions. Most of that discussion has been around how women are often expected to perform emotional labor without compensation both professionally and personally. Women report being asked to perform social glue functions in the workplace without it being part of their job, part of how they are evaluated, or part of how they are paid, and they are culturally expected to perform most of the emotional labor in personal relationships. And perhaps most frustratingly, while men are lauded when they perform emotional labor and mostly given a pass when they don't, the situation is reversed for women who mostly only see punishment for not doing enough.

But ultimately emotional labor is for everyone, and although there is a sex differential in its performance, there is little new I can say on that aspect of the topic. What I can say is something about how emotional labor is related to developmental psychology and cognitive empathy. Specifically, how skill at emotional labor depends on the development of cognitive empathy and lack of cognitive empathy is a limiting factor in being able to perform emotional labor.

I described emotional labor as "managing other people's emotions", but to be more precise emotional labor is acting to influence the emotions of others. To do this one must have some knowledge of the emotions of others and how they can be affected. This knowledge typically comes from either affective empathy or cognitive empathy. Affective empathy is feeling another person's feelings, like being sad because your friend is sad or being scared because a character in a horror movie is scared. Affective empathy's source is probably mirror neurons, and a lack of affective empathy is associated with sociopathy. For this reason affective empathy is sometimes also called "primitive" empathy because it seems to naturally develop on its own and is rarely missing in a person.

Cognitive empathy, on the other hand, is the skill of thinking about others ontologically and is anything but "primitive". In order to be able to think of ways to make your friends happy or worry about what others will think of you, you must model other people and predict their responses. Those models can be simple, like how children employ thing and thing-relationship levels of phenomenological ontological complexity, but such simple models often fail if not backed up by affective empathy. As people age they build up enough cognitive empathy to effectively participate in society without necessarily feeling everyone's feelings, and they develop system level and higher ontological complexity that enables cognitive empathy techniques like seeing other people as made up of parts, distinguishing others' revealed and stated identities, and understanding others' needs and wants. And if a person continues down this path they may develop a generalized sense of cognitive empathy that can tackle broad axiological questions about how to treat themselves and others.

Yet affective empathy and cognitive empathy rarely exist in isolation. In the context of emotional labor, people often first feel—use affective empathy to notice—that an

opportunity exists to affect someone else's emotions, and then use cognitive empathy to figure out what to do. And when cognitive empathy fails us we may fall back on affectively informed actions. This will work most of the time, but pesky philosopher that I am, I want to know what happens in the edge cases, like when you can do something to hurt someone else's feelings without them finding out.

Consider the case of the broken vase. I'm having a fancy dinner party and you lend me your vase to use as a centerpiece. On the way home I stumble and drop the vase, shattering it into a million pieces. Luckily this happens right in front of a store where I can purchase an exact replica, so I immediately replace it. The dinner party goes well, and I "return" the vase with you unable to tell I've replaced it. I have two options:

1. Say nothing about the break.
2. Tell you that I broke the vase and replaced it.

If I have no affective or cognitive empathy it seems likely I will do (1) since it is naively the option that produces the better payout: you'll be mildly happy in (1), whereas there's some chance you'll be angry in (2). If I have no cognitive empathy but plenty of affective empathy, it seems likely I will do (2) because I will feel the bad feelings you would feel if you knew I broke your vase and won't think about the fact that you don't know that I broke it. If I have no affective empathy but plenty of cognitive empathy, though, it now becomes a bit more complex to figure out how I will act. Maybe I will want to spare your feelings and do (1), but maybe I will reason that you would want (2) because it conveys information about me you want to know, and I do it out of a reasoned expectation that acting in this way will more create the world I want to live in. And the situation remains substantially the same if I have plenty of affective empathy to go along with my cognitive empathy, however my feelings will likely affect my axiological calculations in deciding which action to prefer.

In this scenario cognitive empathy enabled emotional labor. Without it I was left either playing a simple game or acting on my feelings with no consideration for you, and so my emotional "labor" was reduced to calculating a payout matrix or dealing with my own conflicting emotions. Cognitive empathy made possible real emotional labor, though, because it provided an ontology to reckon with. True, you might object, it still produced one of the outcomes that could be achieved without cognitive empathy, but emotional labor matters at the margins when we consider many cases where acting without cognitive empathy would produce inconsistent answers.

This is important because without doing enough emotional labor to come to a wise course of action, a desire to help someone borne out of earnest empathy for them may end up unintentionally hurting them. If I failed to understand you sufficiently well in the vase case when I was using cognitive empathy to perform emotional labor, I might have chosen to do (1) when actually (2) is what you would have preferred or vice versa. When we help we risk hurting if we do so unwisely, so helping depends on having the skill to accurately predict how our actions will affect others. This is why people say emotional labor is draining: not only is it mentally challenging but the stress of failure can weight so heavy on us that we find it hard to act.

So what can you do if you want to perform the emotional labor that will allow you to help others as they want to be helped? My own solution is to target virtue when my calculations are insufficiently calibrated, but otherwise you might take the same advice I'll always give: do the emotional labor you fear doing and give up on hoping for the emotional labor you want done for you, because even if you hurt people in the

short run this will enable the necessary personal growth towards increased complexity that will let you help them in the future.

Post Fit Review

My idea here is to have a post where we can discuss thoughts on how well particular posts fit LW without cluttering the post's own comments. StackExchange sites do something similar where reviewers will sometimes post in meta to discuss posts they were unsure if they should perform moderation on, although the focus here is really meant to be on fit (but maybe we'll want to expand the notion later).

Format for comment threads here is:

1. Top level comment is link to post.
2. Comments are thoughts on if the post fits LW or not.

Why I Quit Social Media

This is a linkpost for https://srconstantin.wordpress.com/

Epistemic Status: Personal

I'm not a Puritan. I eat dessert, enjoy a good cocktail, and socialize a lot. I like fun, and I don't think fun is wrong.

So I really understand the allergy to messages of ascetic self-denial.

But lately I've found it necessary to become more like this... and less like this.

More balanced, deliberate, reflective. Less needy, emotionally unstable, dramatic, and attention-seeking.

I've written about this process here, here, here, here, and here. I've been at it for about a year.

Why is it worth being less of a drama queen? In some ways, equinamity is a lot less fun.

But, ultimately, being a drama queen is a dependent's lifestyle. It makes you unable to function on your own. As a practical matter, I am not a dependent, and so I sometimes need to do things — in real life, outside of my own head, where the actual state of the world matters.

I also think it's wrong to constantly interrupt things other people are doing to change the subject to All About Me.

And, basically, that's what social media does. It distances you from reality, makes you focus on a shadow-world of opinions about opinions about opinions; it makes you more impulsive and emotionally unstable; it incentivizes derailing conversations to fish for ego-strokes.

I don't dislike petty bullshit — I enjoy it all too much. I could happily spend eternity picking fights and chasing drama if somehow that were feasible.

So I asked myself "is there, ultimately, anything wrong with living in a world of screams and shadows and impulse pleasures? Do I actually care about anything else?" And the answer was "Unfortunately, yeah. I have to literally sustain my own life, and there are people I genuinely care about. So…ok, reality matters."

And if reality matters, *obviously you shouldn't be doing stuff that makes you into a moron*.

Life after social media isn't hard, in my experience. Life without one pleasure isn't miserable, because there are other pleasures. The brain's pleasure mechanisms are damnably homeostatic; you adjust to about the same amount of average happiness, regardless of how intense or mild the pleasures in your daily life. I miss the drama of social media now and then, but not most days.

I think if you consider yourself reality-oriented or "serious", then quitting social media should be overdetermined.

I’m a little more ambivalent about all that — I’m the kind of person who might plug myself into the Experience Machine — but I think as long as we live on a planet with limited resources, a pure life of fantasy is suicidal, and at least sometimes we have to deal with reality. And we should at least not mislead, or dissipate the efforts of, people who are trying to deal with reality.

Plus, even for dreamers like myself, I think there might someday be a better Annwfn than Facebook.

Value Arbitrage

This is an extremely useful idea that I've learned just recently. In fincance, "arbitrage" means buying and selling things in different markets to take the advantage of difference in price. Like buying a toy in India for $15, and then selling it for $25 in the US.

Turns out, you can arbitrage not just products, but also information, knowledge, skill, or even human relationships.

For example, let's say you have met two amazing people who don't know each other. When you introduce them to each other, you generate a massive amount of value for both of them essentially out of nothing - you have given each of them a gift of knowing someone great.

Another example, is arbitraging information. Eliezer Yudkowsky have learned a lot about rationality and human biases from the books he has read, and then parlayed it into his blog, and later a Harry Potter fanfiction - and that has generated a huge amount of value to the people on the internet, who would never have been exposed to these ideas otherwise.

You can also arbitrage skills. That simply means applying good ideas you have learned in one field to a completely different one. I'm constantly applying the skills and ideas I've learned as a 3D artist to my programming process, and, surprisingly, the ideas I've learned from programming are making me a much better artist.

Finally, an unexpected idea that I have discovered, is that you can arbitrage fiction tropes. You can take a trope you have seen in one story, apply it to a completely different genre, and it will create a new and interesting idea. You can take a SciFi trope and set it in a fantasy world, or you can take a plot structure from a dramatic TV show and use it in your comedy script. In fact, this is how a lot of movie ideas are created.

Smart ideas are often universal. Learn a mental pattern in one field - and apply it elsewhere. Now that I got pretty good at it, every time I have an epiphany in one of my crafts, I intentionally go through each of the other fields I'm interested in and apply it there. This has helped me to generate a lot of very interesting ideas.

Moderator's Dilemma: The Risks of Partial Intervention

If you ever end up moderating a forum, or just becoming deeply involved in the meta section of one like me, it is almost inevitable that you will become involved in disputes over exactly what one is and is not allowed to say on it. You may find these choices can be very difficult as, more often than not, there are no good options. You typically have two choices:

(Note: This post touches on hot-button issues more than I'd like. I'll probably come back later and edit it to avoid these)

- **Libertarian approach to moderation**

In this approach, you refuse to get involved in particular kinds of disputes. The standard example is that you may refuse to moderate posts based on political opinions so long as the poster is polite and they stick to all the other rules. Or you may decide not to moderate insults and flaming, so long as no-one engages in doxxing. This approach has a major downside in that if you aren't moderating based on politics, you'll have to let through posts from white supremacists so long as they are polite and if you aren't moderating insults, you'll get a bunch of jerks using the forum. I'm not saying that this approach is flawed, just that this is what is likely to happen.

- **Non-libertarian approach to moderation**

In this approach, you concede a need to at least occasionally intervene in a particular kind of dispute such as banning the white supremacists or deleting comments that contain racial slurs. It is really, really good to be able to do this. On the other hand, once you intervene in a particular manner, you create the expectation of intervention. Suppose you've banned the white supremacists, but since people expect you to be consistent you end up extending the ban other hateful ideologies too. Quite quickly, you'll find yourself being forced to render a verdict on a whole host of ideologies, not just the clear cut examples. Before you intervened you could refuse take pick sides in certain disputes, but once you've intervened your decision not to intervene in another situation will be taken to mean that you don't consider a particular ideology hateful. If this is a contentious issue, then people will be unhappy no matter how you decide; often you would much prefer to not have to make a ruling on this issue. Again, I'm not saying that this approach is flawed, just that if you choose it, you will most likely run into this issue.

Broader **principle**

Even if you don't moderate a forum, you will see this dilemma crop up in many other situations.

- Suppose you are in charge of a company that is deciding whether or not to have a policy on what its employees post on social media. If you don't have such a policy, they may say horrible things that are offensive and destroy your company's reputation. If you do have such as policy and you choose not to fire someone over something they post, it will be seen as you believing that it is not

hateful. So you may actually end up with a worse reputation than if you didn't have a policy at all.
- Suppose that you on a government panel that is deciding the extent to which you should regulate the safety of toys. You might decide to take a minimal regulatory approach and only ban the most unsafe toys. This would allow to protect a number of people from harm, but you've now also created the expectation that any toy not banned must be safe. You could try writing the opposite on your website or try spending some of your minimal marketing budget informing people, but realistically most people will make this assumption regardless of what you do. You might even end up increasing the total number of injuries by luring consumers into a false sense of security.
- Suppose you write up a set of rules for your club. This makes it easier to ensure that everyone is aware of them. However, in the absence of any written rules the expectation is that you should use your common sense. When these rules are written down, people will start to assume that if something isn't in the rules, it must be allowed, particularly if it would have been easy to put it in the rules if they had wanted to. So you may actually find that more people end up doing the things that you don't want.
- During Coronavirus, some governments have made restrictions like banning gatherings of 100 people or more. This has been taken by people as meaning that it's okay run gatherings with 99 people.

This is scary. In many of these cases, it seems incredibly obvious those in power should at least, even if they do nothing else, regulate the worst cases. But quite quickly we see that this isn't as obvious as it sounds.

The mechanism in the Moderator's Dilemma is that refusing to ban a particular action is seen as either implicitly endorsing it, or at least claiming that it is "not bad". If this isn't the case, then you don't run into the Moderator's Dilemma. For example, you may be able to avoid the Moderator's Dilemma if you have a "Historical Schelling Point". For example, if you ban racial slurs there may be disputes over what counts as a slur, but because there is precedent of treating these differently from other things that are offensive, you may be able avoid having to adjudicate a flamewars where no slurs occur.

Relationship to Other terms

This is related to a few other terms:

- **Precedent** is a term most associated with legal contexts. Courts try to keep the law consistent so that people can follow it, so once one court rules one way, other courts will tend to rule in a manner consistent with this. However, while Precedent is primarily about consistency and only secondarily about how the rules might later expand, the Moderator's Dilemma is primarily about expansion, but also more about being forced to issue a ruling than about the outcomes.
- **Slippery Slopes** are an argument that is often claimed to be a logical fallacy, but which is only poor reasoning if you fail to justify why we are likely to inevitably progress along the slope. The Moderator's Dilemma is a kind of slippery slope, but while the focus of the Slippery Slope is the endpoint and how bad it is, the Moderator's Dilemma is more about being forced to issue a ruling when you don't want to.

Conclusion

I believe that Terminology is Important and so the purpose of this article was to describe a particular situation that I have seen in multiple contexts, but which I did not already have a name for. Group rationality is hard because you can't just look at the immediate effects, but all of the effects that might occur downstream. This is a small attempt to grapple with these problems by identifying one particularly common downstream pattern.

Note: Please try to avoid turning the comments section into a debate on the political examples included in this post. I tried to avoid discussing these topics more than necessary, but these were the clearest examples that I could think of.

Frontpage Posting and Commenting Guidelines

Welcome to the new *LessWrong!* Our goal with the *LessWrong* frontpage is to host high-quality discussion on a wide range of topics, in a way that allows users to make better collective progress toward the truth.

New posts automatically get posted to a user's *personal blog*, where people are free to talk about whatever they like. By default, moderators will consider whether the post is a good fit for the frontpage. If you don't want the post to appear on frontpage, you can uncheck "moderators can promote".

1. Things to shoot for on frontpage

1.1. *Usefulness, novelty, and fun.* The frontpage of this site is for serious intellectual engagement with interesting ideas, with a focus on ideas that are important but challenging to evaluate. Topics that lack inherent importance are OK if the discussion quality is high enough, and particularly if the discussion is useful for other purposes, like building skills; but the best topics will usually be consequential and neglected ones.

1.2. *Accuracy, kindness, and relevance to the discussion at hand*, in the spirit of the Victorian Sufi Buddha ideal.

1.3. *Clarity and openness about what you believe, your reasons for believing it, and what would cause you to change your mind.* Try to make concrete **predictions** and **bets**, and to note the **cruxes** for your beliefs, where possible. It's not always easy to clearly articulate a belief, and it's great to note places where you're uncertain about what you believe, about your reasons, and about your cruxes. We don't want people to feel like they have to conceal or immediately abandon their beliefs whenever those beliefs turn out to be nontrivial to articulate or justify. But incremental progress toward more clarity and openness, even if it's incomplete, is highly valued here.

A corollary of 1.3 is that we often prefer descriptive language (including language describing your current beliefs, emotional state, etc.) over prescriptive language, all else being equal. Prescriptions are obviously an essential part of communication, but descriptions are generally easier to relate to evidence, predictions, and cruxes. We encourage putting a focus on them for that reason.

2. Things to keep to a minimum

2.1. *Community-focused discussion* — i.e., discussion about the *LessWrong*/rationality community, as opposed to discussion about particular object-level topics. We want to avoid dynamics like (from Feynman):

> When I was in high school, one of the first honors I got was to be made a member of the "Arista," which was a group of kids who got good grades, hmm? Everybody wanted to be a member of the Arista, and when I got into this Arista I discovered that what they did in their meetings was to sit around to discuss who else was "worthy" to join "this wonderful group that we are," okay?

If you want to discuss the community more generally, and you don't expect the discussion to be of much interest to people who just want to talk about object-level issues (in psychology, or physics, or zoology, or cryptography, or…), it's best to leave it in your personal blog section.

Questions about the site itself are welcome in the **Meta** section.

2.2. *Crowdedness* **—** i.e., topics that are already really widely discussed in the public sphere, and where it will therefore be harder to say something new.

2.3 *Things of fleeting importance* — i.e., topics that will only be of interest for a couple of weeks, like discussions of what a politician has been doing. We want the frontpage of *LessWrong* to serve both as a training ground for aspiring rationalists and as an archive of accumulated collective knowledge. The ideal discussion will therefore both help build skills and help build knowledge that are valuable down the line. Not every discussion needs to achieve that ideal, but it's a useful one to keep in mind.

We may build features in the future that are for more short-form and clearly ephemeral content on *LessWrong*. If so, this will be in a new section of the site built to be less like a repository of timeless information and discussion, and more like (e.g.) a Facebook feed.

2.4 *Hot-button political issues* **—** Highly politicized issues tend to be very viral, which can often lead to them dominating discussion. These issues often (though not always) score poorly on tractability and neglectedness; they're often emotionally charged in ways that make convergence and skill-building more challenging; and discussion is often triggered by transient news items, as opposed to deep new insights that will be equally relevant years down the line. "Politics is an important domain to which we should individually apply our rationality—but it's a terrible domain in which to learn rationality". This means that highly politicized issues will often score poorly on 1.1, 2.2, and 2.3.

Of course, what counts as a "hot-button political issue" isn't always clear, and we don't want to encourage agonizing or arguing about what counts. (See 2.1.) We just want to encourage users to use their judgment and do their best to keep it to a minimum, so that other topics aren't crowded out.

3. Off-limits things

3.1. *Serious violations of discourse norms* — Threatening behavior, needlessly harsh personal attacks, harassment, doxxing, and so on.

3.2. *Consistently disruptive or low-quality content* — Spam, discussion derailing, and so on.

A list of users with bans or public warnings can be found here.

4. How moderation works

Compared to moderators on other online forums, moderators on *LessWrong* are granted greater ability to change and improve the website, and are trusted with more information. These roles of responsibility are only given to trusted members of the community, and they are known as the *Sunshine Regiment.*

The new, weighted karma system is designed to bring good content to the top. However, this karma system is based on the voting patterns of many individuals, most of whom do not have the time to reflect on big-picture trends, nor the resources to substantially change those trends. In a classic tragedy of the commons, when there are thousands of people voting, no individual is incentivised to spend a lot of time considering their vote.

The incentives set up by the karma system can be considered the community's System 1, and the Sunshine Regiment can be thought of as the community's System 2. Sunshines think about what incentive gradients are being produced, and are given the resources to influence the incentive gradients in a more substantial way (e.g. karma rewards on comments), allowing the community to plan around obstacles and achieve more complex goals.

There are no hard rules about what comments each member of the Sunshine Regiment will give karma rewards to. If your submission has received a karma reward, it will be signified by a small star icon on that comment or post. If your submission has been removed by a Sunshine, they will leave a note explaining why the comment was inappropriate or unsuited to the *LessWrong* frontpage.

Members of the Sunshine Regiment will have access to more information than other users, allowing them to notice negative patterns of behaviours, such as sockpuppet accounts and mass downvoting. The extra information is:

- Access to the identities of voters on any comment/post, and to the voting history of all users.
- The IP address a user wrote a post or comment from.

Sunshines of the 1st LessWrong Regiment are:

- Jim Babcock (jimrandomh)
- Rob Bensinger (RobBensinger)
- Satvik Souza Beri (SatvikBeri)
- Ruby Bloom (Ruby)
- Adom 'Quincy' Hartell (ahartell)
- Elizabeth Van Nostrand (elizabeth)
- Ben Pace (Ben)
- Keller Scholl (Celer)
- Kaj Sotala (Kaj_Sotala)

Related Links

- Moderation Reference (A repository of reasoning and judgments the moderators have used that requires a bit of effort to explain, which we wanted to be able to easily reference if the issue came up again)
- Moderation List of Warnings and Bans
- Posts on philosophy of moderation:
  - Well Kept Gardens Die by Pacifism (Eliezer Yudkowsky)
  - Models of Moderation (Habryka)
  - Public Archipelago (Raemon)

The Best Self-Help Should Be Self-Defeating

Cross-post from blog.

[Self-help is supposed to get people to stop needing it. But typical incentives in any medium mean that it's possible to get people hooked on your content instead. A musing on how the setup for writing self-help differs from typical content.]

Say you have a job as a Self-Help Guru. You spend your days giving out your worldly advice to those who seek guidance on their problems.

I claim that if you're doing job as a Self-Help Guru right, then you should never have repeat customers.

That's the gist behind the idea that the best self-help should be self-defeating.

Here are some analogies of things that I think are like self-help:

- The point of taking antibiotics is so that you eventually stop taking them and feel better.
- The point of wearing glasses is so that you can stop squinting and see more clearly.
- The point of reading a programming textbook is so you can eventually start writing programs on your own.

My claim is that if you're trying to do self-help right, you want people to be able to "graduate" from your ideas and figure out what actually works for them. In a sense, you want to catalyze people to find their own optimal solutions instead of consisting coming back to you for more help every time.

You want them to go off on their own adventures in life, confident that they have the ability to craft new tools when the ones you give them stop working.

Once again, the Recognize vs Generate dichotomy comes into play here. Following advice someone else gave you can look about the same as coming up with something similar on your own. But being the sort of person who can generate solutions independently is far more effective in the long run.

Basically I claim the whole point of self-help is to help people help themselves.

Not that controversial a viewpoint. The real problem, I think, comes in when we consider the way that self-help gets publicized and published.

First off, consider the incentives of many media like newspapers, television channels, novels, or vloggers. Growing an audience is an explicit part of their goals; after all, their profits are largely tied to their viewership. As a result, it makes a lot of sense for them to come out with constant content—it keeps the original crowd coming, and a constant presence means it can draw more people in.

But self-help is different. Arguably, the content isn't even the real point; the real point is something about the self-help "attitude" which enables them to solve future problems. You want them to level up and then head off to do great things in the real world.

Entertainment isn't the point, so you don't really want people binging on your content. You want them to read the content, learn whatever lessons are useful for them, and then move on.

Which means that some sort of constant, fluctuating, or even decreasing number of readers can actually be a sign that you're doing things correctly.

This stark conflict between typical media incentives for publicity and the lofty goals of self-help hits at the heart of the issue, I think. My claim is that basically everyone trying to do self-improvement has gone way off into the "maximize profits and publicity" direction rather than the "maximize beneficial impact of the content" direction.

Here's a hypothetical situation: someone with genuinely altruistic motivations might want to first write some clickbait-type articles to bring in an audience, and then provide more real insights.

"I'll just do the trashy stuff first, and then I'll gradually transition to deeper stuff later," they think.

And that's fair; none of this self-help stuff is happening in a vacuum. Battles for attention in the modern world are zero-sum, and the other side (i.e. all other media) is already optimizing the hell out of "attention-grabbiness".

But that's perhaps too unrealistic. Basically no one is that calculating. I'm not explicitly accusing self-help writers of being evil masterminds who write addictive content under the guise of self-improvement in order to make profits.

Rather, I just think that throughout the normal course of writing self-help content, writers will just have to make certain decisions which trade off the direct benefit of content.

One obvious example is the choice in media format. Books are self-contained and seem to stand at one end of an axis which has Twitter tweets and Facebook posts on the other. For books, feedback from the reader is far less immediate (worse for the author), and the payoff to the reader is far less instant (worse for the reader). But if you go all the way to the other end, you're trading off conceptual complexity, the ability to explain deeper ideas.

There are also subtler things. For example, there are certain design choices when making blogs to reduce binging, like removing infinite scroll, which are probably in your readers' best interests. Most of those choices will also reduce traffic on your site as a whole.

It's not even media format or design choices:

When you think you have good content, you're going to want to share it to others. After all, the only way that self-help materials help others is if people read them in the first place.

And it’s just the case that most ways to get more people interested and spread your content involve increasing the level of “memetically stickiness” or “fun-to-read-ness”, both of which are orthogonal to “ability to level up the reader” and often trade off against it.

The incentive structure for self-help is unfortunately set up in such a way that the traditional ways of cultivating engagement don’t work well with it.

I think one of the worst things that can happen is for people to become a sort of “insight junkie”, where they’re always craving the next mental model or productivity hack, rather than staring down the obvious advice.

The tldr; here is that this is a potential trap for people (me included!) who think they have good content to share. Attempts to reach a wider audience and become more memetically sticky can backfire when you end up getting people hooked on insight-porn-esque longform essays, rather than going out into their lives and being more awesome.

Blind Goaltenders: Unproductive Disagreements

If you're worried about an oncoming problem and discussing it with others to plan, your ideal interlocutor, generally, is someone who agrees with you about the danger. More often, though, you'll be discussing it with people who disagree, at least in part.

The question that inspired this post was "Why are some forms of disagreement so much more frustrating than others?" Why do some disagreements feel like talking to a brick wall, while others are far more productive?

My answer is that some interlocutors are 'blind goaltenders'. They not only disagree about the importance of your problem, they don't seem to understand what it is you're worried about. For example, take AI Safety. I believe that it's a serious problem, most likely the Most Important Problem, and likely to be catastrophic. I can argue about it with someone who's read a fair chunk of LessWrong or Bostrom, and they may disagree, but they will understand. Their disagreement will probably have gears. This argument may not be productive, but it won't be frustrating.

Or I could talk to someone who doesn't understand the complexity of value thesis or orthogonality thesis. Their position may have plenty of nuances, but they are missing a key concept about our disagreement. This argument may be just as civil - or, given my friends in the rationalsphere, *more* civil - but it will be much more frustrating, because they are a blind goaltender with respect to AI safety. If I'm trying to convince them, for example, *not* to support an effort to create an AI via a massive RL model trained on a whole datacenter, they may take into account specific criticisms, but will not be blocking the thing I care about. They can't see the problem I'm worried about, and so they'll be about as effective in forestalling it as a blind goalie.

Things this does not mean

Blind goaltenders are not always wrong. Lifelong atheists are often blind goaltenders with respect to questions of sin, faith, or other religiously-motivated behavior.

Blind goaltenders are not impossible to educate. Most people who understand your pet issue now were blind about it in the past, including you.

Blind goaltenders are not stupid. Much of the problem in AI safety is that there are a great deal of smart people working in ML who are nonetheless blind goaltenders.

Goaltenders who cease to be blind will not always agree with you.

Things this does mean

Part of why AI safety is such a messy fight is that, given the massive impact if the premises are true, it's rare to understand the premises, see all the metaphorical soccer balls flying at you, and still disagree. Or at least, that's how it seems from the perspective of someone who believes that AI safety is critical. (Certainly most people who disagree are missing critical premises.) This makes it very tempting to characterize people who are well-informed but disagree, such as non-AI EAs, as being

blind to some aspect. (Tangentially, a shout-out to Paul Christiano, who I have strong disagreements with in this area but who definitely sees the problems.)

This idea can reconcile two contrasting narratives of the LessWrong community. The first is that it's founded on one guy's ideas and everyone believes his weird ideas. The second is that anyone you ask has a long list of their points of disagreement with Eliezer. I would replace them with the idea that LessWrong established a community which understood and could see some core premises; that AI is hard, that the world is mad, that *nihil supernum*. People in our community disagree, or draw different conclusions, but they understand enough of the implications of those premises to share a foundation.

This relates strongly to the intellectual turing test, and its differences with steelmanning. Someone who can pass the ITT for your position has demonstrated that they understand your position and why you hold it, and therefore are not blind to your premises. Someone who is a blind goaltender can do their best to steelman you, even with honest intentions, but they will not succeed at interpreting you charitably. The ITT is both a diagnostic for blindness and an attempt to cure it; steelmanning is merely a more lossy diagnostic.

The Great Filter isn't magic either

A post suggested by James Miller's presentation at the Existential Risk to Humanity conference in Gothenburg.

Seeing the emptiness of the night sky, we can dwell upon the Fermi paradox: where are all the alien civilizations that simple probability estimates imply we should be seeing?

Especially given the ease of moving within and between galaxies, the cosmic emptiness implies a Great Filter: something that prevents planets from giving birth to star-spanning civilizations. One worrying possibility is the likelihood that advanced civilizations end up destroying themselves before they reach the stars.

The Great Filter as an Outside View

In a sense, the Great Filter can be seen as an ultimate example of the Outside View: we might have all the data and estimation we believe we would ever need from our models, but if those models predict that the galaxy should be teeming with visible life, then it doesn't matter how reliable our models seem: they must be wrong.

In particular, if you fear a late great filter - if you fear that civilizations are likely to destroy themselves - then you should increase your fear, even if "objectively" everything seems to be going all right. After all, presumably the other civilizations that destroyed themselves thought everything seemed to going all right. Then you can adjust your actions using your knowledge of the great filter - but presumably other civilizations also thought of the great filter and adjusted their own actions as well, but that didn't save them, so maybe you need to try something different again or maybe you can do something that breaks the symmetry from the timeless decision theory perspective like send a massive signal to the galaxy...

The Great Filter isn't magic

It can all get very headache-inducing. But, just as the Outside View isn't magic, the Great Filter isn't magic either. If advanced civilizations destroy themselves before becoming space-faring or leaving an imprint on the galaxy, then there is some phenomena that is the cause of this. What can we say, if we look analytically at the great filter argument?

First of all suppose we had three theories - early great filter (technological civilizations are rare), late great filter (technological civilizations destroy themselves before becoming space-faring), or no great filter. Then we look up at the empty skies, and notice no aliens. This rules out the third theory, but leaves the relative probabilities of the other two intact.

Then we can look at objective evidence. Is human technological civilization likely to end in a nuclear war? Possibly, but are the odds in the 99.999% range that would be needed to explain the Fermi Paradox? Every year that has gone by has reduced the likelihood that nuclear war is very very very very likely. So a late Great Filter may seemed quite probable compared with an early one, but much of the evidence we see

is against it (especially if we assume that AI - *which is not a Great Filter!* - might have been developed by now). Million-to-one prior odds can be overcome by merely 20 bits of information.

And what about the argument that we have to assume that prior civilizations would also have known of the Great Filter and thus we need to do more than they would have? In your estimation, is the world currently run by people taking the Great Filter arguments seriously? What is the probability that the world will be run by people that take the Great Filter argument seriously? If this probability is low, we don't need to worry about the recursive aspect; the ideal situation would be if we can achieve:

1. Powerful people taking the Great Filter argument seriously.
2. Evidence that it was hard to make powerful people take the argument seriously.

Of course, successfully achieving 1 is evidence against 2, but the Great Filter doesn't work by magic. If it looks like we achieved something really hard, then that's some evidence that it is hard. Every time we find something unlikely with a late Great Filter, that shifts some of the probability mass away from the late great filter and into alternative hypotheses (early Great Filter, zoo hypothesis,...).

Variance and error of xrisk estimates

But let's focus narrowly on the probability of the late Great Filter.

Current estimates for the risk of nuclear war are uncertain, but let's arbitrarily assume that the risk is 10% (overall, not per year). Suppose one of two papers comes out:

1. Paper A shows that current estimates of nuclear war have not accounted for a lot of key facts; when these facts are added in, the risk of nuclear war drops to 5%.
2. Paper B is a massive model of international relationships with a ton of data and excellent predictors and multiple lines of evidence, all pointing towards the real risk being 20%.

What would either paper mean from the Great Filter perspective? Well, counter-intuitively, papers like A typically increase the probability for nuclear war being a Great Filter, while papers like B decrease it. This is because none of 5%, 10%, and 20% are large enough to account for the Great Filter, which requires probabilities in the 99.99% style. And, though paper A decreases the probability of the nuclear war, it also leaves more room for uncertainties - we've seen that a lot of key facts were missing in previous papers, so it's plausible that there are key facts still missing from this one. On the other hand, though paper B increases the probability, it makes it unlikely that the probability will be raised any further.

So if we fear the great filter, we should not look at risks whose probabilities are high, but risks who's uncertainty is high, where the probability of us making an error is high. If we consider our future probability estimates as a random variable, then the one whose variance is higher is the one to fear. So a late Great Filter would make biotech risks even worse (current estimates of risk are poor) while not really changing asteroid impact risks (current estimates of risk are good).

Motivating a Semantics of Logical Counterfactuals

(**Disclaimer:** *This post was written as part of the CFAR/MIRI AI Summer Fellows Program, and as a result is a vocalisation of my own thought process rather than a concrete or novel proposal. However, it is an area of research I shall be pursuing in the coming months, and so ought to lead to the germination of more ideas later down the line. Credit goes to Abram Demski for inspiring this particular post, and to Scott Garrabrant for encouraging the AISFP participants to actually post something.*)

When reading Soares' and Levinstein's recent publication on decision theory, *Death in Damascus*, I was struck by one particular remark:

> Unfortunately for us, there is as yet no full theory of counterlogicals [...], and for [functional decision theory (FDT)] to be successful, a more worked out theory is necessary, unless some other specification is forthcoming.

For some background: an FDT agent must consider counterfactuals about what *would* happen if its decision algorithm on a given input were to output a particular action. If the algorithm is deterministic (or the action under consideration is merely outside of the range of possible outputs of the agent), then it is a logical impossibility that the algorithm produces a different output. Hence the initial relevance of logical counterfactuals or counterlogicals: counterfactuals whose antecedents represent a *logical* impossibility.

My immediate thought about how to tackle the problem of finding a full theory of counterlogicals was to find the right logical *system* that captures counterlogical inference in a way adequate to our demands. For example, Soares and Fallenstein show in *Toward Idealized Decision Theory* that the principle of explosion (from a contradiction, anything can be proved) leads to problematic results for an FDT solution. Why not simply posit that our decision theory uses paraconsistent logic in order to block some inferences?

Instead, and to my surprise, Soares and Levinstein appear to be more concerned about finding a *semantics* for logical counterfactuals - loosely speaking, they are looking for a uniform interpretation of such statements which shows us what they really *mean*. This is what I infer from reading the papers it references on theories of counterlogicals such as Bjerring's *Counterpossibles*, which tries to extend Lewis' possible-worlds semantics for counterfactuals to cases of logical counterfactuals.

The approaches Soares and Levinstein make reference to do not suffice for their purposes. However, this is not because they give proof-theoretic considerations a wide berth, as I thought previously; I now believe that there is something that the semantic approach does get *right*.

Suppose that we found some paraconsistent logical system that permitted precisely the counterlogical inferences that agreed with our intuition: would this theory of logical counterfactuals satisfy the demands of a fully-specified functional decision theory? I would argue not. Specifically, this approach seems to give us a merely a posteriori, ad hoc justification for our theory - given that we are working towards an *idealised* decision theory, we ought to demand that the theory that supports it is *fully motivated*. In this case, this cashes out as making sure our counterlogical statements

and inferences has a *meaning* - a meaning we can resort to to settle the truth-value of these counterlogicals.

This is not to say that investigating different logical systems is entirely futile: indeed, if the consideration above were true and a paraconsistent logic could fit the bill, then a uniform semantics for the system would serve as the last piece of the puzzle.

In the next year, I would like to investigate the problem of finding a full theory of logical counterfactuals, such that it may become a tool to be applied to a functional decision theory. This will, of course, involve finding the logical system that best captures our own reasoning about logical counterfactuals. Nonetheless, I will now also seek to find an actual motivation for whichever system seems the most promising, and the best way to find this motivation will be through finding an adequate semantics. I would welcome any suggestions in the comments about where to start looking for such a theory, or if any avenues have thus far proved to be more or less promising.

Why Attitudes Matter

Sometimes when I am giving ethical advice to people I say things like "it's important to think of yourself and your partner as being on the same team" or "just remember that women in short skirts are almost certainly not wearing short skirts to arouse you in particular" or "cultivate your curiosity and desire to know what's actually going on."

I get pushback on this. After all, I am a consequentialist. Why am I talking about people's attitudes instead of their actions? It doesn't matter what I think of the woman in the short skirt, as long as I refrain from being a dick to her because of her clothing choices.

An emphasis on attitudes can be really bad for some people. Some people, having been given the advice that they should cultivate their curiosity, will spend a lot of time navel-gazing about whether they're really curious and whether this curiosity counts as curiosity and maybe they are self-deceiving and actually just want to prove themselves right Not only is this really unpleasant, but if you're spending all your time navel-gazing about whether you're sufficiently curious you're never actually going to go buy a book about the Abbasid empire. It completely fails to achieve the original goal. If this is a problem you're prone to, I think my attitude-based advice is probably not going to be helpful, although I can't give any other advice; I personally get as much navel-gazing as I can stand trying to keep my obviously shitty attitudes in check, and don't have any introspective energy left over for anything else.

Nevertheless, I think an attitude emphasis can be really important, for two reasons.

First, for any remotely complicated situation, it would be impossible to completely list out all the things which are okay or not okay. For instance, think about turning my "think of yourself and your partner as being on the same team" advice into a series of actions. You might say "it is wrong to insult your partner during disagreements." But for some people, insults are part of resolving disagreements. Saying "I am not sure you've really thought this through" rather than "that is the stupidest fucking idea I've ever heard" feels artificial to them, like they're walking on eggshells. For them, intimacy requires the ability to say exactly what you're feeling, without softening it.

Or you might say "if you think of arguments for your partner's side, then say it." However, this might lead you to fall victim to what C S Lewis in the Screwtape Letters called the Generous Conflict Illusion:

> Later on you can venture on what may be called the Generous Conflict Illusion. This game is best played with more than two players, in a family with grown-up children for example. Something quite trivial, like having tea in the garden, is proposed. One member takes care to make it quite clear (though not in so many words) that he would rather not but is, of course, prepared to do so out of “Unselfishness”. The others instantly withdraw their proposal, ostensibly through their “Unselfishness”, but really because they don’t want to be used as a sort of lay figure on which the first speaker practices petty altruisms. But he is not going to be done out of his debauch of Unselfishness either. He insists on doing “what the others want”. They insist on doing what he wants. Passions are roused. Soon someone is saying “Very well then, I won’t have any tea at all!”, and a real quarrel ensues with bitter resentment on both sides. You see how it is done? If each side had been frankly contending for its own real wish, they would all have kept within

> the bounds of reason and courtesy; but just because the contention is reversed and each side is fighting the other side’s battle, all the bitterness which really flows from thwarted self-righteousness and obstinacy and the accumulated grudges of the last ten years is concealed from them by the nominal or official “Unselfishness” of what they are doing or, at least, held to be excused by it. Each side is, indeed, quite alive to the cheap quality of the adversary’s Unselfishness and of the false position into which he is trying to force them; but each manages to feel blameless and ill-used itself, with no more dishonesty than comes natural to a human.

Or you might say "if your partner seems to be making a mistake, give them some friendly advice, without being overly critical." But some people are naturally controlling-- not abusive, just the sort of people who get upset when their partner loads the dishes a different way than they're used to or prefers to read a map rather than using the GPS. Those people might very well decide that they shouldn't give any friendly advice, for much the same reason that an alcoholic shouldn't go to a bar. It never stops after one.

If you are thinking about the situations from a position of "my partner and I are both on Team Our Collective Happiness And Well-Being," then the answer to all these thorny situations becomes clear. You should give a word of friendly advice, unless you are the sort of person who is incapable of stopping at a word of friendly advice. You should speak in a way that makes your partner and you feel more intimate and able to resolve conflicts, rather than less so. You should say "hm, I think the vacation you want to go on is cheaper" but you should not do the Generous Conflict Illusion. And so on and so forth.

Second, an attitude emphasis prevents rules-lawyering. Whenever you list a set of actions, there are a certain number of people who will figure out how to get as close as possible to breaking the rules, and then complain when you get annoyed at them, because technically they didn't break any rules. (Rules-lawyering is particularly likely to happen in issues of sexual ethics, but it is certainly not reserved for those situations.) For example, they might say "you said I wasn't supposed to yell at my wife or call her nasty names! You never specifically said I wasn't supposed to respond to my wife forgetting to do the dishes by piling up all the dirty dishes onto her bed."

But obviously if you are two people cooperating to solve the problem of the dirty dishes piling up, "stick the dishes on the other person's bed" is not how you would respond. (Unless, I guess, they agreed ahead of time that this was a useful if disgusting way to help them remember-- like I said, it's really hard to make hard-and-fast rules.) That is a way you'd respond if you're approaching the situation as a war between you and your partner, and the winner is whoever gets a clean sink while having to do the least dishes. This is, to put it lightly, not a good way of solving your relationship problems.

I suspect that action-based advice works best in relatively simple situations where there aren't a lot of possible actions and where there are few situations that require a judgment call: for instance, it works great for "don't hit people unless they started it". Attitude-based advice works best for complicated situations where there are lots of possible ways of fucking up: for instance, it works well for intimate relationships, intellectual or artistic life, and career choice.

Musings on Double Crux (and "Productive Disagreement")

Epistemic Status: Thinking out loud, not necessarily endorsed, more of a brainstorm and hopefully discussion-prompt.

Double Crux has been making the rounds lately (mostly on Facebook but I hope for this to change). It seems like the technique has failed to take root as well as it should. What's up with that?

(If you aren't yet familiar with Double Crux I recommend checking out Duncan's post on it in full. There's a lot of nuance that might be missed with a simple description.)

Observations So Far

- Double Crux hasn't percolated *beyond* circles directly adjacent to CFAR (it seems to be learned mostly be word of mouth). This might be evidence that it's too confusing or nuanced a concept to teach without word of mouth and lots of examples. It might be evidence that we have not yet taught it very well.
- "Double Crux" seems to refer to two things: the specific action of "finding the crux(es) you both agree the debate hinges on" and "the overall pattern of behavior surrounding using Official Doublecrux Technique". (I'll be using the phrase "productive disagreement" to refer to the second, broader usage)

Double Crux seems hard to practice, for a few reasons.

Filtering Effects

- In local meetups where rationality-folk attempt to practice productive disagreement on purpose, they often have trouble finding things to disagree about. Instead they either:
    - are already filtered to have similar beliefs,
    - quickly realize their beliefs shouldn't be that strong (i.e. they disagree on Open Borders, but soon as they start talking they admit that neither of them really have that strong an opinion)
    - they have wildly different intuitions about deep moral sentiments that are hard to make headway on in a reasonable amount of time - often untethered to anything empirical. (i.e. what's more important? Preventing suffering? Material Freedom? Accomplishing interesting things?)

Insufficient Shared Trust

- Meanwhile in many online spaces, people disagree all the time. And even if they're both nominally rationalists, they have an (arguably justified) distrust of people on the internet who don't seem to be arguing in good faith. So there isn't enough foundation to do a productive disagreement at all.
- One failure mode of Double Crux is when people disagree on what frame to even be using to evaluate truth, in which case the debate recurses all the way to the level of basic epistemology. It often doesn't seem to be worth the effort to resolve that.
- Perhaps most frustratingly: it seems to me that there are many longstanding disagreements between *people who should totally be able to communicate*

clearly, update rationally, and make useful progress together, and those disagreements don't go away, people just eventually start ignoring each other or leaving the dispute as unresolved. (An example I feel safe bringing up publicly is the argument between Hanson and Yudkowsky, although this may be a case of the 'what frame are we even using' issue above.)

That last point is one of the biggest motivators of this post. If the people I most respect can't productively disagree in a way that leads to *clear progress, recognizable from both sides*, then what is the rationality community even doing? (Whether you consider the primary goal to "raise the sanity waterline" or "build a small intellectual community that can solve particular hard problems", this bodes poorly).

Possible Pre-Requisites for Progress

There's a large number of sub-skills you need to productively disagree. To have *public norms* surrounding disagreement, you not only need individuals to have those skills - they need to trust that each other have those skills as well.

Here's a rough list of those skills. (Note: this is long, and it's less important that you read the whole list than that the list is long, which is why Double Cruxing is hard)

- **Background beliefs** (listed in Duncan's original post)
    - Epistemic humility ("I could be the wrong person here")
    - Good Faith ("I trust the other person to be believing things that make sense to them, which I'd have ended up believing if I were exposed to the same stimuli, and that they are generally trying to find the the truth")
    - Confidence in the existence of objective truth
    - Curiosity / Desire to uncover truth
- **Building-Block and Meta Skills**
- (Necessary or at least very helpful to learn everything else)
    - Ability to gain habits (see Trigger Action Plans, Reflex/Routines, Habits 101)
    - Ability to introspect and notice your internal states (Focusing and Noticing can help)
    - Ability to induce a mental state or reframe
    - Habit of gaining habits
- **Notice you are in a failure mode, and step out.** Examples:
    - You are fighting to make sure an side/argument wins
    - You are fighting to make another side/argument lose (potentially jumping on something that seems *allied* to something/someone you consider bad/dangerous)
    - You are incentivized to believe something, or not to notice something, because of social or financial rewards,
    - You're incentivized not to notice something or think it's important because it'd be physically inconvenient/annoying
    - You are offended/angered/defensive/agitated
    - You're afraid you'll lose something important if you lose a belief (possibly 'bucket errors')
    - You're rounding a person's statement off to the nearest stereotype instead of trying to actually understand and response to what they're saying
    - You're arguing about definitions of words instead of ideas
    - Notice "freudian slip" ish things that hint that you're thinking about something in an unhelpful way. (for example, while writing this, I typed out

"your opponent" to refer to the person you're Double Cruxing with, which is a holdover from treating it like an adversarial debate)

(The "Step Out" part can be pretty hard and would be a long series of blogposts, but hopefully this at least gets across the ideas to shoot for)

- **Social Skills** (i.e. not feeding into negative spirals, noticing what emotional state or patterns other people are in [*without* accidentaly rounding them off to a stereotype])
    - Ability to tactfully disagree in a way that arouses curiosity rather than defensiveness
    - Leaving your colleague a line of retreat (i.e. not making them lose face if they change their mind)
    - Socially reward people who change their mind (in general, frequently, so that your colleague trusts that you'll do so for them)
    - Ability to listen (in a way that makes someone feel listened to) so they feel like they got to actually talk, which makes them inclined to listen as well
    - Ability to notice if someone else seems to be in one of the above failure modes (and then, ability to point it out gently)
    - Cultivate empathy and curiosity about other people so the other social skills come more naturally, and so that even if you don't expect them to be right, you can see them as helpful to at least understand their reasoning (fleshing out your model of how other people might think)
    - Ability to communicate in (and to listen to) a variety of styles of conversation, "code switching", learning another person's jargon or explaining yours without getting frustrated
    - Habit asking clarifying questions, that help your partner find the Crux of their beliefs.
- Actually Thinking About Things
    - Understanding when and how to apply math, statistics, etc
    - Practice thinking causally
    - Practice various creativity related things that help you brainstorm ideas, notice implications of things, etc
    - Operationalize vague beliefs into concrete predictions
- **Actually Changing Your Mind**
    - Notice when you are confused or surprised and treat this as a red flag that something about your models is wrong (either you have the wrong model or no model)
    - Ability to identify what the actual Crux of your beliefs are.
    - Ability to track bits of small bits of evidence that are accumulating. If enough bits of evidence have accumulated that you should at least be taking an idea *seriously* (even if not changing your mind yet), go through motions of thinking through what the implications WOULD be, to help future updates happen more easily.
    - If enough evidence has accumulated that you should change your mind about a thing... like, actually do that. See the list of failure modes above that may prevent this. (That said, if you have a vague nagging sense that something isn't right even if you can't articulate it, try to focus on that and flesh it out rather than trying to steamroll over it)
    - Explore Implications: When you change your mind on a thing, don't just acknowledge, actually think about what other concepts in your worldview should change. Do this
        - because it *should* have other implications, and it's useful to know what they are....

        - because it'll help you actually retain the update (instead of letting it slide away when it becomes socially/politically/emotionally/physically inconvenient to believe it, or just forgetting)
    - If you notice your emotions are not in line with what you now believe the truth to be (in a system-2 level), figure out why that is.
- **Noticing Disagreement and Confusion, and then *putting in the work* to resolve it**
- If you have all the above skills, and your partner does too, and you both trust that this is the case, you can still fail to make progress if you don't actually follow up, and schedule the time to talk through the issues thoroughly. For deep disagreement this can take years. It may or may not be worth it. But if there are longstanding disagreements that continuously cause strife, it may be worthwhile.

Building Towards Shared Norms

When smart, insightful people disagree, at least one of them is doing something wrong, and it seems like we should be trying harder to notice and resolve it.

A rough sketch of a norm I'd like to see.

Trigger: You've gotten into a heated dispute where at least one person feels the other is arguing in bad faith (especially in public/online settings)

Action: Before arguing further:

- stop to figure out if the argument is even worth it
- if so, each person runs through some basic checks (i.e. "am *I* being overly tribal/emotional?)
- instead of continuing to argue in public where there's a lot more pressure to not lose face, or steer social norms, they continue the discussion privately, in whatever the most human-centric way is practical.
- they talk until *at least* they succeed at Step 1 Double Crux (i.e. agree on where they disagree, and *hopefully* figure out a possible empirical test for it). Ideally, they also come to as much agreement as they can.
- Regardless of how far they get, they write up a short post (maybe just a paragraph, maybe longer depending on context) on what they *did* end up agreeing on or figuring out. (The post should be something they both sign off on)

Epistemic Spot Check: Exercise for Mood and Anxiety (Michael W. Otto, Jasper A.J. Smits)

Introduction

Everyone knows exercise (along with diet and sleep) makes a big difference in depression and anxiety. Depressed and anxious people are almost by definition bad at transforming information about how to improve their lives into actions with large up front costs, so this data is not as useful as it might be. Exercise for Mood and Anxiety (Michael W. Otto, Jasper A.J. Smits) aims to close that gap by making the conventional wisdom actionable. It does that through the following steps:

1. Present evidence that exercise is very helpful and why, to create motivation.
2. Walk you through setting up an environment where exercise requires relatively little will power to start.
3. Scripts and advice to make exercise as unmiserable as possible while you are doing it.
4. Scripts and advice to milk as much mood benefit as possible from a given amount of exercise.
5. An idiotic chapter on weight and food.

Parts 3 and 4 use a lot of techniques from cognitive behavioral therapy and mindfulness, and I suspect there's a second order benefit of learning to apply these techniques to a relatively easy thing, so you can apply them to the rest of your life later.

Epistemic Spot Checking

Claim: "a study of 55,000 adults in the United States and Canada found that people who exercised had fewer symptoms of anxiety and depression." (Kindle Locations 103-104).

Correctly cited, paper has no proof of causation. (abstract) (PDF) The study does in fact say this, but it also says "Despite the fact that none of these surveys [of which this paper is a metaanalysis] was [sic] originally designed to explore this association… ". I'm not saying you can never repurpose data, but with something like this where the real question is causality, it seems suspicious. The authors do consider the idea that causation runs from mental health (=energy, hopefulness, executive function) -> exercise and dismiss if, for reasons I find inadequate.

Claim: "Other studies add to this list of mood benefits by indicating that exercise is also linked to less anger and cynical distrust, as well as to stronger feelings of social integration." (Kindle Locations 104-106).

Correctly cited, paper has no proof of causation. (Abstract).

Claim: And these benefits don't just include reducing symptoms of distress in people who have not been formally diagnosed with depression or anxiety. The benefits of exercise also include lower rates of psychiatric disorders; there is less major depression, as well as fewer anxiety disorders in those who exercise regularly. (Kindle Locations 107-109).

Correctly cited, paper has no proof of causation.

The dismissal of causality goes on for another three citations but I'm just going to skip to the intervention studies. Otto gives these population studies more credence than I would but does note that the intervention studies are more informative.

Claim: study summarized 70 studies on this topic and showed that adults who experience sad or depressed moods, but not at levels that meet criteria for a psychiatric disorder, reliably report meaningful improvements in their mood as they start exercising. (Kindle Locations 116-117).

Correctly cited, study accuracy undetermined. (Full paper). My fear (based on spot checking a similar book you'll see in the rejects post) is that each of these studies consists of 15 people. All the metaanalysis in the world won't save you if you do 100 small studies and only publish the 50 that say what you want. The studies included go all the way back to 1969: I can't decide if that makes them more informative or less.

Claim: The latest estimates are that about 17% of adults experience a major depressive episode in their lifetimes and that about half who have it experience recurrent episodes over time. (Kindle Locations 124-126).

True. (Full paper). The same study is cited for both facts, but I can only find the 50% statistic in the paper. The data is kind of old (started in 1981), but of course you can't get 30-year data except by starting 30 years ago. This paper says the lifetime prevalence of mood disorders (depression, bipolar 1 and 2, and their baby siblings) is 20%; this study puts prevalence in the US at 16.9%.

Claim: As is the case with major depressive disorder, anxiety disorders are common, affecting more than 1 in 4 (28.8%) adults in their lifetimes" (Kindle Locations 136-137).

True. (Full paper). He cites the same paper I did for the 20% mood disorder statistic.

Claim: [Anxiety disorders] tend to be especially long-lasting when people do not receive treatment. (Kindle Locations 137-138).

True, although not particularly specific. (Full paper)

Claim: Exercise in itself is a stressor—it requires effort, and it forces the body to adapt to the demands placed on it. (Kindle Locations 141-142).

True. (Full paper).

Claim: A study examined firefighters reaction to stress, and then gave half a 16 week exercise course. The study group showed improvements in stress responses. (Kindle location 148)

True. (Abstract) (PDF). I really like this study. The group presumably had a high baseline fitness level, so this isn't the difference between couch potato and a walk. And they have before and after metrics. The study is marred only by the small sample size (53).

Claim: "stress plays a key role in both the development and the continuation of depression and anxiety disorders." (Kindle Locations 152-153).

Accurate citation, very complicated topic. (Abstract).

Okay, it is becoming clear I don't have the time to check every one of these citations and you don't have time to read it. From here on out please assume a baseline of very dense citations, all of which accurately report the study results, if with a little more confidence than the study design merits, and I'm only going to call out things that deserve special attention on account of controversy or importance.

Claim: exercise increases serotonin just like the primary class of anti-depressants, selective serotonin update inhibitors.

True but less relevant than implied. They're relying on a model of how SSRIs treat depression that is fairly outdated. SSRIs definitely increase serotonin, it's just that there's no evidence that's their mechanism of action against depression except that they do it and they treat depression. "Depression is caused by a serotonin deficiency" is a lie simplification told to patients and their families to allay fear and shame around psychiatric treatment. This doesn't undercut their point that exercise is good for you, but does indicate this is not a great book to learn brain chemistry from.

Claim: Both aerobic (prolonged moderate exercise such as running, cycling, or rowing over time) and anaerobic (like weight lifting or short sprinting) exercise have been found to be effective for decreasing depression, (Kindle Locations 239-241).

True. (Study 1 PDF) (Study 2 abstract).

Empirical Results

The theory behind this book is very well supported; the prescriptions it makes flow naturally from the theory, but the authors present no direct evidence that they work. I'm torn about this. I don't want to engage in RCT worship; having a systemic understanding of a problem is even better than evidence a particular solution worked better or worse than another solution in a different population. On the other hand, humans are very complicated and it's easy to identify the problem but guess the wrong solution.

I couldn't test any of this on myself because I already enjoy exercise for a lot of reasons, so I scrounged up an unscientific sample from my wider social network to try it.

14 people filled out the pre-book survey. 3 people filled out the post-attempt survey. None of them exercised more.

Summary

The theory sections of this book are my high water mark for scientific rigor in a self-help-psych book. I'm currently reading a lot of those with the goal of finding out how much rigor is reasonable to expect, so that's high praise.

The book walks the very fine line between reassuring and condescending, which is pretty unavoidable with CBT and mindfulness.

I did not like the last chapter and recommend skipping it. It feels like they tried to stuff all the usual diet-and-exercise stuff in at the end. Some of my problem is I think their recommendations are wrong, and some is that I believe that even if they were correct, throwing them in at the last minute undercuts the message of the book.

The first part of this is that, in America, at least in certain subcultures, any mention of weight makes the whole thing About Weight. Too many people use health or mood as a socially acceptable way to say "you're not hot enough", so any mention of weight in the context of diet or exercise automatically makes weight the real topic of the conversation. If the improvements in mood are enough of a reason to exercise, let them be enough, and the weight loss can be a pleasant surprise or not happen, and both are okay because you got what you came for.

The authors compound this problem by using Body Mass Index as a guide for goal weight. BMI is completely unsuited for use in individuals, even more so for people who just started gaining muscle mass. If you must talk about fat in the context of health use body fat percentage or certain circumference ratios (e.g. wrist:stomach).

The second problem is the speed with which EFMaA tries to address nutrition. The book (correctly) treats exercise as a thing that is challenging to start despite all its benefits, and spends 10 chapters explaining why it's worth trying and providing scripts to make it workable for you, for the sole benefit of mood, ignoring everything else you might get out of exercise. I don't know why the authors thought that that required an entire book but the even more complicated of nutrition for every possible benefit of nutrition could be squeezed into half a chapter. I would be have been very excited for another book by the same authors about how to implement healthy eating, but the half assed treatment here makes me pause.

They also present a particular diet as the settled science, when there is no such thing in nutrition. "Eat produce and fish" is fairly uncontroversial, but they recommend a lot more refined grains than many other people. I don't know who is correct, but it was disappointing to see a book that had been so rigorous up to that point blithely paint over controversy.

[I have emailed Michael Otto about the handling of nutrition and have yet to hear back].

Speaking of which *Exercise for Mood and Anxiety* mentions that both aerobic (cardio) and anaeorbic (weights) are good for mood, but every single example is cardio, with an occasional cardio + core strength.

Mixed in through the book are tales of how Olympic athletes motivate themselves. This feels spectacularly irrelevant to me. I don't want to win a gold medal, I want to climb V2s and be happy.

You might find this book valuable if:

- You want some ideas (although not conclusive proof) around how exercise helps mood.
- You want to want to exercise, and want scripts and tools to transform that into “want to exercise right now.”
- You find exercise unpleasant and want to get the best trade of unpleasantness-for-benefits possible.
- You would like to treat a mood issue with exercise (whether it reaches the level of official disorder or not).
- You want to change how you think about exercise (for improving your mood or something else).
- You are interested in CBT or mindfulness and want to practice with the large print version before tackling them directly.
- You think you are different than my test audience.

You probably won’t find this book valuable if:

- You already have an exercise program you are happy with.
- You have body image or eating disorder issues (last chapter only, and a single section of the 10th, the rest of it is fine).
- You want prescriptions for a particular exercise program, as opposed to general principles.
- You want to learn the nitty gritty of how exercise affects mood.
- You are similar to my test audience.

Post supported by Patreon.

Thinking on the page

“Thinking on the page” is a handle that I’ve found useful in improving my writing (and my introspection more generally). When I write, for the most part, I’m trying to put something that I already feel is true into words. But when I think on the page, the words are getting ahead of my internal sense of what’s true. I’m writing something that just sounds good, or even just something that logically flows from what I’ve already written but has come untethered from my internal sense. It’s kind of a generalized verbal overshadowing.

I don’t think this is challenging only to people who think [of themselves as thinking] non-verbally, considering how much more universal are experiences like “this puts exactly what I believe into words better than I ever could” or even the satisfaction of finding a word on the tip of the tongue. Some people seem to be better than others not just at describing their internal sense of truth, but at tapping into it at all. But if you think only in internal monologue, you may have a very different perspective on “thinking on the page”—I’d be interested to hear about it.

At best, this is what happens in what Terry Tao calls the “rigorous” stage of mathematics education, writing, “The point of rigour is not to destroy all intuition; instead, it should be used to destroy bad intuition while clarifying and elevating good intuition.” At worst, it’s argument based on wordplay. Thinking on the page can be vital when you’re working beyond your intuition, but it’s equally vital to notice that you’re doing it. If what you’re writing doesn’t correspond to your internal sense of what’s true, is that because you’re using your words wrong, or because you need to use the page as prosthetic working memory to build a picture that can inform your internal sense?

The two places this becomes clearest for me are in academic writing and in art critique. Jargon has the effect of constantly pulling me back towards the page. If it doesn’t describe a native concept, I can either heroically translate my entire sense of things and arguments about them into jargon, or I can translate the bare raw materials and then manipulate them on the page—so much easier. As for art, the raw material of the work is already there in front of me—so tempting to take what’s easy to point to and sketch meaning from it, while ignoring my experience of the work, let alone what the raw material had to do with that experience.

A lack of examples often goes hand in hand with thinking on the page. Just look at that last paragraph: “translate”, “raw materials”, “manipulate”—what am I even talking about? An example of both the jargon and art failure modes might be my essay about Yu-Gi-Oh! Duel Monsters. My analysis isn’t entirely a joke, but it’s not a realistic reading in terms of the show’s experiential or rhetorical effect on the audience, intended or otherwise. The protagonist’s belief in the heart of the cards and his belief in his friends are genuinely thematically linked, but neither one is the kind of “shaping reality by creative utterance” that has anything to do with how the characters talk their way around the procedural weirdness of the in-show card game as game. But when I put all these things in the same language, I can draw those connections just fine. I’m playing a game with symbols like “creative utterance”.

How can one notice when this is happening? Some clues for me:

- I feel like I’m moving from sentence to sentence rather than back and forth between thought and sentence
- I feel something like “that’s not quite right”
- I feel a persistent “tip of the tongue” sensation even after writing
- I feel clever
- I haven’t used an example in a while
- I’m using jargon or metaphor heavily

What can one do after noticing?

- Try to pull the words back into your internal picture, to check whether they fit or not—they might, and then you’ve learned something
- Rewrite without self-editing until something feels right, with permission to use as many words or circumlocutions as it takes
- Try to jar the wording you want mentally into place by trying more diverse inputs or contexts (look at distracting media, related essays, a thesaurus)
- Ask “but is that true?”
- Connect with specific examples
- Focus on the nonverbal feeling you want to express; try to ask it questions

What’s a good way to practice?

- Write reviews of art/media you encounter, then read other people’s reviews. As far as “not being led astray by thinking on the page” is more than the zeroth-level skill of writing-as-generically-putting-things-into-words, I think this is a good place to practice what’s particular to it. People seem to have a good enough sense of what they liked and why for good criticism to resonate, but often not enough to articulate that for themselves, at least without practice. So it can be good to pay attention to where the attempted articulations go wrong.
- Write/read mathematical proofs or textbook physics problems, paying attention to how the steps correspond to your sense of why the thing is true (or using the steps to create a sense of why it’s true)
- If it seems like the sort of thing that would do something for you, find or develop a meditation practice that involves working with "felt senses" (I don't have a real recommendation here, but it's the kind of thing Focusing aims for)

The goal isn’t to eliminate thinking on the page, but to be more deliberate about doing it or not. It can be useful, even if I haven’t said as much about that.

One thing I don’t recommend is using “you’re thinking on the page” as an argument against someone else. If you find yourself thinking that, it’s probably an opportunity to try harder to get in their head. Like most of these things, as a way thinking can go wrong, this is a concept best wielded introspectively.

(Here’s a puzzle: if this is a first-level skill, can you go up another level? If I’m saying something like “felt senses/thoughts want to make words available”, then what things “want to make felt senses available”? Can you do anything with them?)

[cross-posted from my personal blog]

The Five Hindrances to Doing the Thing

> "*Compare your normal level of consciousness with that of an athlete in the zone, or with a person in an emergency. You'll realize that daily life consists mostly of different degrees of dullness and mindlessness.*"
>
> - Culadasa (John Yates, Ph.D.), *The Mind Illuminated*

It has been noted that the term *akrasia* does not seemingly carve reality at the joints in a useful way. The general problem of knowing that you should do a thing and yet having trouble getting yourself to do it is an ancient problem, and luckily, people have actually been working on solving it for a long time.

A useful approach for solving a *general* problem is often to thoroughly solve a *specific instance* of that problem and then to try to generalize it. For thousands of years, humans have been working on solving a specific very difficult problem: how to make themselves sit down for up to an hour a day and meditate.

Meditation, on its face, is an overwhelmingly boring task, providing almost no intrinsic reward, especially for beginners. It is almost the degenerate case of making yourself do the most boring possible thing within the realm of actual human activity. It might (or might not) come as a surprise that over the centuries, those who teach meditation have narrowed the potential obstacles to meditation to only five: Desire, aversion, laziness or lethargy, agitation due to worry or remorse, and doubt. In this article, I will attempt to generalize these five hindrances to be applicable to any task, not just meditation. In so doing, I hope to provide a road map to fighting akrasia in the moment.

The following section is organized as follows: Each hindrance has specific qualities, which will help you recognize its occurrence; a cause, which explains why this specific form of resistance arises; and remedies to be employed in the moment of recognizing it.

1. Desire
    - Qualities
        - Distraction due to intrusive thoughts of pleasures related to material existence or attempts to avoid their opposites.
        - Gain/loss, fame/obscurity, pleasure/pain, praise/blame.
    - Cause
        - Desire is the sensation of wanting to obtain or keep something, and is healthy and good in the appropriate context.
    - Remedies
        - "Unification of mind" - cognitively emphasize the utility of having a mind that is singularly engaged on the current task, and set the firm

intention to do so.

        - Briefly address the negative consequences that would come along with getting whatever pleasure it is you feel distracted by.
        - Focus on the pleasure of the *current moment*. If you are actively engaged with an important task, there will be pleasurable sensations and thoughts associated with that fact. Give yourself a mental pat on the back. Let yourself appreciate your own attentiveness and specifically note the quality of pleasure associated with doing so.

2. Aversion
    - Qualities
        - Resistance, rejection, denial, dissatisfaction, judgement, self-accusation and boredom.
        - Wanting things to not be the way they are.
        - Fear is a case of being averse to something that hasn't occurred.
        - Pain causes aversion, but isn't itself aversion.
    - Cause
        - Aversion motivates us to avoid or eliminate what is unpleasant or harmful.
    - Remedies
        - Rest and narrow your focus toward the task at hand. Increase your concentration.
        - Conversely, you can broaden your focus to your whole sensorium, making the distracting thought/sensations seem diminished by comparison, and then refocus your attention on the task at hand.
        - Similarly to the remedy for Desire, focus on the pleasure/happiness to be found in the present moment, and whatever other present-oriented positive mental states can be recognized.
        - If your aversive thoughts involve ill-will toward other people or yourself, switch briefly to producing feelings of good-will for your target, then gently put your attention back on the task at hand.
3. Laziness or lethargy
    - Qualities
        - Procrastination, sleepiness, lack of motivation.
        - Laziness arises when the cost of an activity seems to outweigh the benefits.

        - Lethargy arises when there seems to be nothing interesting or rewarding going on.
    - Causes
        - Laziness motivates us to conserve our energy for tasks that might be more valuable.
    - Remedies
        - Intentionally muster up the motivation for the task by focusing on future rewards.
        - "Just do it" - plunge into the task despite resistance and focus on the positive qualities of the task (i.e. just start the task and if Aversion arises, then employ the remedies for Aversion).
        - Try to remain in the moment, focus on what you're actually doing, stop watching the clock.
        - Do more. Go faster. Engage harder. If the task isn't stimulating you, push the throttle until it does.
        - If you suspect your torpor to be of a partly physical origin, re-invigorate your body and mind by moving around, drinking some water, splashing cold water on your face, then aggressively engage with the task again.
4. Agitation due to worry and remorse
    - Qualities
        - Agitation due to possible consequences.
        - Anxiety due to imagined scenarios.
    - Causes
        - Helps us prepare for an uncertain future.
    - Remedies
        - Resolve to take positive action to correct past mistakes.
        - Cognitively let go of past mistakes.
        - Intentionally focus on the present joy and pleasantness of the moment. Joyfulness both makes it difficult to focus on negative events or possibilities, and situates remorse in a more psychologically tolerable context, from which you can move past it in the moment.
        - As with Aversion, you can either broaden your focus to the entire current context, or narrow it back onto the specific thing you need to do. One of these will probably feel more "right". Agitation has its own

Aversive quality, so it makes sense that some of the remedies for Aversion will work here, and vice versa.

5. Doubt
    - Qualities
        - Focus on negative results or outcomes.
        - A positive mental quality that becomes pathological when one focuses only on the emotional quality which saps motivation rather than performing a cognitive appraisal of the utility of the task.
        - May involve comparing your own performance to others'.
    - Causes
        - Keeps us from wasting our time and resources on unnecessary activities.
        - Invites us to question our current behavior with reasoned skepticism.
    - Remedies
        - Use reasoning abilities to fully engage with and dissolve the doubt (or find it to be a valid doubt, if rational analysis proves it to be so).
        - Doubt is dispelled by the trust and confidence that comes from success; success comes from persistent effort; just keep going!
        - Re-examine your motivation and ensure it is powerful and convincing.
        - Sustained attention to the task at hand.

Bonus meta-skills:

- When you find your mind wandering from the task at hand, rejoice. Give yourself a big smile and mental cheer. Imagine that sound that plays when your character gains a level. If you react this way, you'll condition yourself to notice mind-wandering. Do not beat yourself up. If you yell at yourself, you'll condition yourself to *avoid noticing* mind-wandering.
- Likewise, when you notice reluctance to start or continue a task, engage with the resistance with curiosity and objectivity, and try to lightly and playfully ferret out which of the above hindrances might be in play. Use the acronym RAIN: Recognize, Accept, Investigate, Non-Identification. This means you notice the resistance, you accept its presence neutrally, investigate it calmly, and don't identify with whatever you find. Give yourself a pat on the back when you employ this algorithm. The last thing you need is aversion about your aversion.
- You may have noticed that some variety of "engage with the present moment and focus on the intrinsic pleasurable qualities of whatever it is you're doing" appear as remedies for more than one of the hindrances. You can preempt the manifestation of most of the hindrances by maintaining this kind of mindful, present-oriented, enjoyment-based approach to your tasks. There's something

engaging and pleasurable to be found in any possible task, even if that task is sitting with your eyes closed.

The five hindrances are likely familiar to you, and I personally find that "akrasia" is just one of these issues surrounded by an aversive haze that obscures its true nature. The remedies may feel obvious to you. Regardless, you'll find it's convenient to have a handy, semi-proven, distilled flowchart of solutions to the specific issues that arise while battling the many-headed hydra of akrasia.

Unfair outcomes from fair tests

This is a linkpost for http://whaaales.com/unfair-outcomes-from-fair-tests/

[Summary: Say you use a fair test to predict a quality for which other non-tested factors matter, and then you make a decision based on this prediction. Then people who do worse on the test measure (but not necessarily the other factors) are subject to different error rates, even if you estimate their qualities just as well. If that's already obvious, great; I additionally try to present the notion of fairness that lets one stop at "the test is fair; all is as it should be" as a somewhat arbitrary line to draw with respect to a broader class of notions of statistical fairness.] [Not sure if this text will appear anywhere in LW 2.0, so this is a test.]

Solomonoff Induction explained via dialog.

This is a linkpost for https://arbital.com/p/solomonoff_induction/?l=1hh

Map of the AI Safety Community

This is a linkpost for https://aisafety.com/wp-content/uploads/2017/09/AI_Safety_Community_Map_Version__1_0.jpg

I have made a map of the AI Safety Community!

The map is greatly inspired by the map of the rationalist community made by Scott Alexander.

There are bound to be omissions and misunderstandings, and I will be grateful for any corrections. I promise that I will incorporate the feedback into a new version of the map.

The sizes of the cities/dwellings reflect my understanding of how much they contribute to AI Safety. The locations and borders reflect my judgement of who focus on what, and I had to make some difficult choices.

(Made with Fractal Mapper 8, and crossposted to AISafety.com and r/controlProblem)

I hope that you will find the map useful, and find inspiration to visit new places.