Best of LessWrong: February 2018

1. "Cheat to Win": Engineering Positive Social Feedback
2. The Intelligent Social Web
3. The Principled Intelligence Hypothesis
4. Rationality Feed: Last Month's Best Posts
5. Robustness to Scale
6. Why we want unbiased learning processes
7. Replacing expensive costly signals
8. Crypto autopsy reply
9. Inconvenience Is Qualitatively Bad
10. Hufflepuff Cynicism on Crocker's Rule
11. Mapping the Archipelago
12. Arguments about fast takeoff
13. Write a Thousand Roads to Rome
14. Toward a New Technical Explanation of Technical Explanation
15. "Just Suffer Until It Passes"
16. Self-regulation of safety in AI research
17. Circling
18. Introduction to Noematology
19. Walkthrough of 'Formalizing Convergent Instrumental Goals'
20. A Proper Scoring Rule for Confidence Intervals
21. Mythic Mode
22. Two types of mathematician
23. Two Coordination Styles
24. The Monthly Newsletter as Thinking Tool
25. Beware Social Coping Strategies
26. Beware arguments from possibility
27. Factorio, Accelerando, Empathizing with Empires and Moderate Takeoffs
28. The abruptness of nuclear weapons
29. An alternative way to browse LessWrong 2.0
30. Rationalist Lent
31. Pain, fear, sex, and higher order preferences
32. On Building Theories of History
33. Science like a chef
34. Lessons from the Cold War on Information Hazards: Why Internal Communication is Critical
35. The Logic of Science: 2.2
36. Open-Source Monasticism
37. The Utility of Human Atoms for the Paperclip Maximizer
38. Active vs Passive Distraction
39. Knowledge is Freedom
40. How I see knowledge aggregation
41. Bug Hunt 2
42. The map has gears. They don't always turn.
43. Some conceptual highlights from “Disjunctive Scenarios of Catastrophic AI Risk”
44. Whose reasoning can you rely on when your own is faulty?
45. Hammertime Intermission and Open Thread
46. Confidence Confusion
47. Don't Condition on no Catastrophes
48. Pseudo-Rationality
49. Status: Map and Territory
50. June 2012: 0/33 Turing Award winners predict computers beating humans at go within next 10 years.

Best of LessWrong: February 2018

1. "Cheat to Win": Engineering Positive Social Feedback
2. The Intelligent Social Web
3. The Principled Intelligence Hypothesis
4. Rationality Feed: Last Month's Best Posts
5. Robustness to Scale
6. Why we want unbiased learning processes
7. Replacing expensive costly signals
8. Crypto autopsy reply
9. Inconvenience Is Qualitatively Bad
10. Hufflepuff Cynicism on Crocker's Rule
11. Mapping the Archipelago
12. Arguments about fast takeoff
13. Write a Thousand Roads to Rome
14. Toward a New Technical Explanation of Technical Explanation
15. "Just Suffer Until It Passes"
16. Self-regulation of safety in AI research
17. Circling
18. Introduction to Noematology
19. Walkthrough of 'Formalizing Convergent Instrumental Goals'
20. A Proper Scoring Rule for Confidence Intervals
21. Mythic Mode
22. Two types of mathematician
23. Two Coordination Styles
24. The Monthly Newsletter as Thinking Tool
25. Beware Social Coping Strategies
26. Beware arguments from possibility
27. Factorio, Accelerando, Empathizing with Empires and Moderate Takeoffs
28. The abruptness of nuclear weapons
29. An alternative way to browse LessWrong 2.0
30. Rationalist Lent
31. Pain, fear, sex, and higher order preferences
32. On Building Theories of History
33. Science like a chef
34. Lessons from the Cold War on Information Hazards: Why Internal Communication is Critical
35. The Logic of Science: 2.2
36. Open-Source Monasticism
37. The Utility of Human Atoms for the Paperclip Maximizer
38. Active vs Passive Distraction
39. Knowledge is Freedom
40. How I see knowledge aggregation
41. Bug Hunt 2
42. The map has gears. They don't always turn.
43. Some conceptual highlights from “Disjunctive Scenarios of Catastrophic AI Risk”
44. Whose reasoning can you rely on when your own is faulty?
45. Hammertime Intermission and Open Thread
46. Confidence Confusion
47. Don't Condition on no Catastrophes
48. Pseudo-Rationality

49. Status: Map and Territory
50. June 2012: 0/33 Turing Award winners predict computers beating humans at go within next 10 years.

"Cheat to Win": Engineering Positive Social Feedback

This post outlines a very simple strategy that's been working for me lately. It may be obvious to some, but it only clicked for me recently.

Positive social stimulation is fun for humans, right? We like to be liked. It makes us cheerful. We're motivated to do things that make people smile at us and praise us.

But purely optimizing for being liked is a bad idea for lots of reasons: it leads away from your real goals and values, it motivates you to be deceptive, it's kind of shallow and unsatisfying in the long run.

So here's what you do instead: first, decide what you actually want to do. Then, *seek out people who will socially reward you for doing that, and set yourself up to get social rewards* .

Marketing experts will tell you that you have to "find your tribe", find the fans of your product, and focus on delighting them. It's fine if you have haters. Haters are almost irrelevant. You succeed if you have enough *fans* who value your stuff highly enough.

This applies across areas of life. You only need (about) one job. You only need one spouse. You only need a small number of close friends. Having great supporters is more important than avoiding having any haters.

I used to have the intuition that "fairness" meant I wasn't allowed to bias my social environment in my favor; that I should expose myself equally to people who liked and disliked me, people who did and didn't share my values, in order to get a "balanced" impression of the world.

This is pretty stupid, actually.

You, as a very small creature moving through infinite space, don't learn about the universe by drawing uniform samples from it. You learn through pursuing goals, which means you'll spend more attention on areas of the universe that are *useful to you*, which means things that are easy for you or helpful for your life, things that give you energy and resources to explore more.

An amoeba, as it crawls around, is going to learn more about the parts of the petri dish with food than the parts without. This is because the amoeba is *alive*. So are you.

As a motivational hack towards any kind of project, it really helps to *set yourself up to have recurrent social interactions with people who support you in that project.*

Meetup groups are good for this. Mixers. Mailing lists. Actually select for people who *like* the thing you're into, and it's astonishing how much it'll feel like the "world" supports you!

Use moments when you're in an energetic, upbeat mood to set up plans for things that'll give you positive social feedback in the future -- make plans to meet people or go to events, or apply to things or submit your work to things. That way you get a recurring stream of "good news" in your inbox, which will trigger more upbeat moods

in future. Engineer your social environment to reinforce you for pursuing your goal and you'll be more likely to keep going.

A mastermind group is maybe the most explicit example of this kind of engineering. Get 3-7 people together who have similar goals (starting businesses is a common example) and meet regularly to offer support and cheer on each other's progress. The vibe of the mastermind should be "we're all awesome and we're going to succeed together." It's designed to help you keep up momentum.

Doing this isn't about wireheading or fooling yourself, it's about focusing your attention, including your social attention, in the areas that can offer rewards instead of the barren spaces.

So much defensiveness is *unnecessary*. Unproductive. It's silly to feel like you have to steel yourself against an unfriendly world if you haven't even checked to look for friends. If you take the attitude of "X is cool and awesome -- who's with me on this?" there's a good chance you'll find a community of X-fans. I have seen (and made) *so* many strategic social errors based on the premise that you have to defend yourself against haters rather than seek out fans. It's much better to aim to win than to not-lose.

The Intelligent Social Web

Epistemic status: Fake Framework

When you walk into an improv scene, you usually have no idea what role you're playing. All you have is some initial prompt — something like:

> *"You three are in a garden. The scene has to involve a stuffed bear somehow. Go!"*

So now you're looking to the other people there. Then someone jumps forward and adds to the scene: "Oh, *there* it is! I'm glad we finally found it!" Now you know a *little* bit about your character, and about the character of the person who spoke, but not enough to fully define anyone's role.

You can then expand the scene by adding something: "It's about time! We're almost late now." Now you've specified more about what's going on, who you are, and who the other players are. But it's still the case that *none of you knows what's going on*.

In fact, if you *think* you know, you'll often quickly be proven wrong. Maybe you imagine in that scene you're an uptight punctual person. And then the third person in the scene says to you, "What do you care, Alex? You're always late to everything anyway!" Surprise! Now you need to flush who you thought you were from your mind, accept the new frame, and run with it as part of your newly evolving identity. Otherwise the scene sort of crashes.

It would go more smoothly if you didn't hold any preconceptions about who you are or what's going on. The scene tends to work better if you stay in the present moment and just jump in with the first thing that comes to mind (as long as it's shaped by what has happened so far). Then the collection of interactions and emerging roles spontaneously guides your behavior, which in turn help guide others' behavior, all of which recursively defines the "who" and "what" of the scene. Your job as a player isn't to play a character; it's to *co-create a scene*.

We can sort of pretend that there's a "director": it's the intelligence that emerges between the players via their interactions. It's a distributed system that computes relationships and context by guiding each node in its network to act freely within constraints. From this vantage point, the network guides players, and the job of each player is to be guidable but not purely passive (since a passive node is just relaying information rather than aiding in the computation). As long as everyone involved is plugged into and responsive to this network, the scene will usually play out well.

I suspect that improv works because we're doing something a lot like it pretty much all the time. The web of social relationships we're embedded in helps define our roles as it forms and includes us. And that same web, as the distributed "director" of the "scene", guides us in what we do.

A lot of (but not all) people get a strong hit of this when they go back to visit their family. If you move away and then make new friends and sort of become a new person (!), you might at first think this is just who you are now. But then you visit your

parents… and suddenly you feel and act a lot like you did before you moved away. You might even try to hold onto this “new you” with them… and they might respond to what they see as strange behavior by trying to nudge you into acting “normal”: ignoring surprising things you say, changing the topic to something familiar, starting an old fight, etc.

In most cases, I don’t think this is malice. It’s just that they *need the scene to work*. They don’t know how to interact with this “new you”, so they tug on their connection with you to pull you back into a role they recognize. If that fails, then they have to redefine who *they* are in relation to *you* — which often (but not always) happens eventually.

I’m basically taking as an axiom of this framework that people need the “scene” to work — which is to say, they need to be able to play out their roles in relation to others’ roles within a coherent context. I don’t think *why* this is the case is relevant for using this framework… but I’ll wave my hands at a vague just-so story anyway for the sake of pumping intuition: human beings’ main survival strategy seems to be based on coordinating in often complex ways in tribes. For the individual, this means that *fitting in* becomes paramount. For the group, this means *knowing what to expect from each person* is critical. So a trade becomes possible: the individual can fit into and benefit from the group as long as they’re playing a role that fits well with the collective.

This can result in some pretty strange roles. From this vantage point, a person who repeatedly leaves one abusive relationship only to get into another roughly similar one actually makes a lot of sense: this is a role that this person knows how to play. It’s *horrible*, but it’s still better than not fitting into the social scene. It creates a coherent relationship with someone who’s willing to (or has to) play an “abuser” role, and often with people in “rescuer” roles too. The trap they’re in isn’t (just) that their current abusive partner is gaslighting or threatening them; it’s that they *don’t have another role they can see how to play*. Unless and until that person finds a different one that fits into the social web, the strands of that web will tug them back into their old role. They don’t have enough slack in the web around them to change their fate.

The same kind of web/slack dynamics show up in more pleasant-to-play roles too. The privilege of a middle-class American white man by default has him playing out some kind of roughly known story-like path (probably involving college and having kids and maybe a divorce) that, in the end, will probably still leave him being one of the richest people on Earth. And all the while, he might well have no clue that he has other options or even that he’s on a path — but he’ll still know, somehow, not to step off that path (“I *have to* go to college; are you *crazy*?”). Never mind that his lack of slack here is awfully convenient for him.

I’ve watched religious conversions and deconversions happen via basically the same mechanism. I knew a fellow many years ago (unattached to this community) who was a proud atheist. Then he started dating a Christian girl. Something like a month later, he started quoting the Bible — but “only because they’re handy metaphors” and not because he really *believed* any of that stuff, you see. It later turned out he’d been going to church with her. He kept offering reasons that seemed *vaguely plausible* (“It’s a neat group of people, and it matters to her, and I can take the time to read”), but there’s a *pattern* here that was obvious. A few months later he told me he’d converted. Last I heard they had moved to Utah.

The great part is, I knew this was going to happen when they started dating. Why? Because when I warned him that he might find himself wanting to believe her religion once they started having sex, his reaction was to reassure me by *acting confident that he was immune to this*. That meant he was more focused on managing my perception of him than he was in noticing how the social web was tugging him toward a transition of roles. I didn't know if they'd stay together, but I was pretty sure that if they did, he'd convert.

I could give literally hundreds of examples like this. From where I'm standing, it looks like one of the great challenges of rationality is that people change their minds about meaningful things mostly only when the web tugs them into a new role. *Actually thinking* in a way that for real changes your mind in ways that defy your web-given role is socially deviant, and therefore personally dangerous, and therefore something you're motivated not to learn how to do.

Ah, but if we're immersed in a culture where status and belonging are tied to changing our minds, and we can signal that we're open to updating our beliefs, then we're good… as long as we know Goodhart's Demon isn't lurking in the shadows of our minds here. But surely it's okay, right? After all, we're smart and we know Bayesian math, and we *care* about truth! What could possibly go wrong?

Another challenge here is that the part of us that feels like it's thinking and talking is (usually) analogous to a character in an improv scene. The *players* know they're in a scene, but the *characters* they're playing don't. The characters also aren't surprised about who or what they are: the not-knowing of identity and context is something only the players experience, to open themselves up to the guidance of the distributed "director". This means that (a) the characters are actively wrong about why they do what they do, (b) they are deeply confused about how much sense everything makes, and (c) *they don't know they're confused*.

I claim that most of us, most of the time, are playing out characters as defined by the surrounding web — and we usually haven't a clue how to Look at this fact, much less intentionally use our web slack to change our stories.

I think this is also part of why improv is challenging: you have to set aside the character you would normally play in order to create room for something new.

The web as a whole wants to know what kind of role you're playing, and how well you're going to play it, so that it can know what to expect of you. So, a lot of its distributed resources go into computing a model of you.

One of the more obvious transmission methods is chat — idle gossip, storytelling, speculation, small talk. People sync up their impressions of someone they've met, and try to make sense of surprising events in conversation. If a lover brings their partner some flowers and the recipient freaks out and runs off, suddenly there's a *need to understand*, and the flower-giver might try asking a mutual friend for some help understanding. And even if they do come to understand ("Oh, that's because their last partner brought them flowers to break up with them"), there's often an impulse to *share the story with friends*, so that the web as a whole can hold everyone in sensible roles and make the scene work. ("Oh, we had a funny misunderstanding earlier, poor Sam….")

A lot of this is transmitted more subtly too, in body language and facial expressions and vocal tone and so on. If Bob is “creepy” (i.e., is playing a “creepy” role in the web), then it speaks volumes if everyone who meets Bob then cringes just a tiny bit when he’s later mentioned even if they say only good things about him. This means that someone who has *never met Bob* can get a “vibe” about him from multiple people in a way that shapes how they interpret what Bob says and does when they finally *do* meet him.

Sometimes, some people with enough web-savvy weaponize this. It doesn’t *mean* anything for someone to “be creepy” except that they have a web-like impact on others — which is to say, they have a “creepy” role. In a healthy network, this correlates with something actually meaningfully bad that’s worth tracking. But because perceived roles shape what people expect of a person, it’s enough for a *rumor* to echo through the web in order for someone to be *interpreted* as “creepy”. So a sufficiently cunning person could actually cause someone to be slowly isolated and distrusted without there being any facts at all to justify this as the social web's stance.

(And yes, I’ve seen this happen. Many times.)

The same kind of thing can happen with “positive” labels, too. What it *means* for someone to be fit for a leadership role, in the social web's eyes, is that they are seen as compatible with that role. So if someone is tall, attractive, and either vicious or strong depending on how you choose to see it, it might be enough to have the “strong” interpretation echo more powerfully than the “vicious” one in order for the web to conspire to put them in a leadership position.

…which means that even people who are seen as good leaders might not, in fact, be good leaders in the sense of *making good leadership decisions*. But they are by *definition* good leaders in the sense of *playing the role well*. After all, if the general consensus is that Abraham Lincoln was a great President, then there’s a sense in which *that makes it true*, since that’s what “great” *means* here. The “explanations” thereafter are often stories to justify one’s holding of a popular opinion.

The same thing holds for when someone seems “rational”. This is one reason to worry deeply when members of subgroups internally agree with each other on who is a top-notch clear thinker or “really a rationalist” but disagree with people in other subgroups. This looks less to me like people seeking truth, and a lot more like groups engaging in a subtle memetic battle over what “rational” gets to mean.

From where I’m standing, it looks to me like we’re all immersed in not-knowing, while our “characters” keep talking as though they know what’s going on, implicitly following some hidden-to-them script.

The web encodes a lot of its guidance about what we should expect and how to behave via the structure of stories. Or rather, story structures are what expectations about roles and scenes *are*.

The trouble is, a lot of the stories we *talk about* have the structure of what our characters are supposed to say rather than of what actually happens. Imagine a movie where the new kid at a school gets bullied by the popular kids and then makes friends with quirky outcasts. What happens to the bullies in the end? In real life, bullies often don’t get their comeuppance — but *having this fictional story in our*

hearts lets us play out vivid indignation through our characters in the real-life version. Because the bullies aren't *supposed* to get away with it, right? That wouldn't be *fair*!

Some parts of our story-like intuitions are scripts for what should actually happen. Some are things our scripts say we should think or feel or talk about within our given roles. Some are merely incidental details. Sussing out which parts are which is part of the trick of getting this framework to work for you.

For instance, the stereotypical story of the worried nagging wife confronting the emotionally distant husband as he comes home really late from work… is actually a pretty good caricature of a script that lots of couples play out, as long as you know to ignore the gender and class assumptions embedded in it.

But it's *hard* to sort this out without just enacting our scripts. The version of you that would be thinking about it is your *character*, which (in this framework) can accurately understand its own role only if it has enough slack to become genre-savvy *within the web*; otherwise it just keeps playing out its role.

In the husband/wife script mentioned above, there's a tendency for the "wife" to get excited when "she" learns about the relationship script, because it looks to "her" like it suggests how to *save the relationship* — which is "her" enacting "her" role. This often aggravates the fears of the "husband", causing "him" to pull away and act dismissive of the script's relevance (which is "his" role), driving "her" to insist that they just need to *talk* about this… which is the same pattern they were in before. They *try* to become genre-savvy, but there (usually) just isn't enough slack between them. So their effort merely changes the topic while they play out their usual scene.

So if you don't like the story you're in, how do you *really* change it?

Well, it depends on which "you" is asking the question.

Characters often want change *as part of their role*. And just as importantly, their role often requires that they *can't achieve* that change. The tension between craving and deprivation gives birth to the character's dramatic *raison d'être*. The "wife" can't be as clingy and anxious if the "husband" opens up, so "she" enacts behavior that "she" *knows* will make "him" close down. "She" can't really choose to change this because "her" thwarted desire for change *is part of* "*her*" *role*.

Intentionally creating real personal change requires the *player* to decide to shake things up. Characters avoid understanding this clearly for basically the same reason that most works of fiction avoid breaking the fourth wall.

But I claim there's a way to sidestep this and inject meaningful genre-savviness into your character if you (the player) so choose.

The essence of this is to *stop*.

Just stop.

For a little while, pause the incessant activity, the trying to figure out, the jumping into reaction when a feeling or idea bursts into awareness, the fidgeting to dispel social or physical discomfort instead of savoring it.

Just let all avenues for acting out a role come to stillness.

And then in your stillness, listen closely to your experience as though this is the first moment you've ever experienced anything at all.

This whole process is practically *guaranteed* to make your character flip out. I don't claim you'll *like* doing this (at least at first), or that it'll make sense to you (at first). Maybe it sounds too much like meditation and you have a storm of thoughts associated with that. Maybe the *idea* is so atrocious or ill-founded or mystic-flavored to you that you don't want to even *try* it.

And that's fine! Maybe it makes sense for you to wait until death forces this stillness on you.

But if you choose to try it *anyway*, you can watch as your character does these theatrics…

…and clearly see for yourself that they really are *just theatrics*…

…and you can start to consciously remember who you are beyond all that reactivity.

That reactivity is what takes up slack. When you attend to deep stillness this way, you can directly see for yourself how to create slack, just as clearly as you can feel your tongue in your mouth. And just as clearly, you can watch the ebb and flow of the social web and the ways in which you and everyone else pretends to be bound by its laws.

Then it'll be immensely obvious to you how to create real change in your life.

This was long, so I'll try to summarize:

- You can choose to see social groups at all scales as running a distributed computation across the social web. You can choose to view that process as generating an agent — the intelligent social web — who tries to predict and guide each person's behavior.
- The social web offers each person a trade: prioritize making the scene work, and you'll be included in it. In fact, the web *is* the aggregate efforts of all the people who have accepted that trade. And basically everyone we know about accepts this trade.
- Everything about yourself that you have conscious access to is subject to your role as part of the social web. If you try to defy this, then your fate will play out through your defiance.
- Room for interpretation in your role in the scene means your script has room to change. This is slack in the social web.
- There's a way of directly seeing how to change your fate by Looking, if you so choose. This amounts to something like pausing long enough to clearly see the reactions that try to keep you from pausing.

I'll close this post by noting that there's a meta-level to track here. In the story The Emperor's New Clothes, the child's utterance wasn't enough on its own to pop the illusion:

> "*But the Emperor has nothing at all on!*" *said a little child.*
>
> "*Listen to the voice of innocence!*" *exclaimed his father; and what the child had said was whispered from one to another.*

> "*But he has nothing at all on!*" *at last cried out all the people. The Emperor was vexed, for he knew that the people were right; but he thought the procession must go on now! And the lords of the bedchamber took greater pains than ever, to appear holding up a train, although, in reality, there was no train to hold.*

What if the father had instead responded “No, child, you’re just too foolish to see his fine garments”? He might have, out of fear of what those who were standing nearby might think of him and his kid. Then the child’s simple voice of reason would not be heard.

Or what if the people near the father/child pair had felt too uneasy to pass along what the child had said?

What if the Emperor could have instilled this kind of nervousness in his people ahead of time? He might have thought that there will be innocent children in the parade, and it might have occurred to some part of him that they had best not be taken seriously — to spare others their embarrassment, of course. Then, oh then what strange propaganda they all would see.

Some of the scripts the social web assigns work less well if they’re known. Because of this, the web will often move to silence people who threaten to speak those fragile truths. This can show up, for instance, as people trying to dismiss and discredit *the person saying the idea* rather than just the idea. The arguments usually sound sensible on the surface, but the underlying tone ringing through the strands of the web is “Don’t listen to this one.”

If it’s not clear why I’m mentioning this, then I imagine it’ll become really obvious quite soon.

The Principled Intelligence Hypothesis

I have been reading the thought provoking *Elephant in the Brain*, and will probably have more to say on it later. But if I understand correctly, a dominant theory of how humans came to be so smart is that they have been in an endless cat and mouse game with themselves, making norms and punishing violations on the one hand, and cleverly cheating their own norms and excusing themselves on the other (the 'Social Brain Hypothesis' or 'Machiavellian Intelligence Hypothesis'). Intelligence purportedly evolved to get ourselves off the hook, and our ability to construct rocket ships and proofs about large prime numbers are just a lucky side product.

As a person who is both unusually smart, and who spent the last half hour wishing the seatbelt sign would go off so they could permissibly use the restroom, I feel like there is some tension between this theory and reality. I'm not the only unusually smart person who hates breaking rules, who wishes there were more rules telling them what to do, who incessantly makes up rules for themselves, who intentionally steers clear of borderline cases because it would be so annoying to think about, and who wishes the nominal rules were policed predictably and actually reflected expected behavior. This is a whole stereotype of person.

But if intelligence evolved for the prime purpose of evading rules, shouldn't the smartest people be best at navigating rule evasion? Or at least reliably non-terrible at it? Shouldn't they be the most delighted to find themselves in situations where the rules were ambiguous and the real situation didn't match the claimed rules? Shouldn't the people who are best at making rocket ships and proofs also be the best at making excuses and calculatedly risky norm-violations? Why is there this stereotype that the more you can make rocket ships, the more likely you are to break down crying if the social rules about when and how you are allowed to make rocket ships are ambiguous?

It could be that these nerds are rare, yet salient for some reason. Maybe such people are funny, not representative. Maybe the smartest people are actually savvy. I'm told that there is at least a positive correlation between social skills and other intellectual skills.

I offer a different theory. If the human brain grew out of an endless cat and mouse game, what if the thing we traditionally think of as 'intelligence' grew out of being the cat, not the mouse?

The skill it takes to apply abstract theories across a range of domains and to notice places where reality doesn't fit sounds very much like policing norms, not breaking them. The love of consistency that fuels unifying theories sounds a lot like the one that insists on fair application of laws, and social codes that can apply in every circumstance. Math is basically just the construction of a bunch of rules, and then endless speculation about what they imply. A major object of science is even called discovering 'the laws of nature'.

Rules need to generalize across a lot of situations—you will have a terrible time as rule-enforcer if you see every situation as having new, ad-hoc appropriate behavior. We wouldn't even call this having a 'rule'. But more to the point, when people bring you their excuses, if your rule doesn't already imply an immovable position on every case you have never imagined, then you are open to accepting excuses. So you need

to see the one law manifest everywhere. I posit that technical intelligence comes from the drive to make these generalizations, not the drive to thwart them.

On this theory, probably some other aspects of human skill are for evading norms. For instance, perhaps social or emotional intelligence (I hear these are things, but will not pretend to know much about them). If norm-policing and norm-evading are somewhat different activities, we might expect to have at least two systems that are engorged by this endless struggle.

I think this would solve another problem: if we came to have intelligence for cheating each other, it is unclear why general intelligence *per se* is is the answer to this, but not to other problems we have ever had as animals. Why did we get mental skills this time rather than earlier? Like that time we were competing over eating all the plants, or escaping predators better than our cousins? This isn't the only time that a species was in fierce competition against themselves for something. In fact that has been happening forever. Why didn't we develop intelligence to compete against each other for food, back when we lived in the sea? If the theory is just 'there was strong competitive pressure for something that will help us win, so out came intelligence', I think there is a lot left unexplained. Especially since the thing we most want to explain is the spaceship stuff, that on this theory is a random side effect anyway. (Note: I may be misunderstanding the usual theory, as a result of knowing almost nothing about it.)

I think this Principled Intelligence Hypothesis does better. Tracking general principles and spotting deviations from them is close to what scientific intelligence is, so if we were competing to do this (against people seeking to thwart us) it would make sense that we ended up with good theory-generalizing and deviation-spotting engines.

On the other hand, I think there are several reasons to doubt this theory, or details to resolve. For instance, while we are being unnecessarily norm-abiding and going with anecdotal evidence, I think I am actually pretty great at making up excuses, if I do say so. And I feel like this rests on is the same skill as 'analogize one thing to another' (my being here to hide from a party could just as well be interpreted as my being here to look for the drinks, much as the economy could also be interpreted as a kind of nervous system), which seems like it is quite similar to the skill of making up scientific theories (these five observations being true is much like theory X applying in general), though arguably not the skill of making up scientific theories well. So this is evidence against smart people being bad at norm evasion in general, and against norm evasion being a different kind of skill to norm enforcement, which is about generalizing across circumstances.

Some other outside view evidence against this theory's correctness is that my friends all think it is wrong, and I know nothing about the relevant literature. I think it could also do with some inside view details – for instance, how exactly does any creature ever benefit from enforcing norms well? Isn't it a bit of a tragedy of the commons? If norm evasion and norm policing skills vary in a population of agents, what happens over time? But I thought I'd tell you my rough thoughts, before I set this aside and fail to look into any of those details for the indefinite future.

Rationality Feed: Last Month's Best Posts

I write a daily rational feed. I write up summaries/teasers for the previous day's article that I found interesting and/or enjoyable. I follow most rationalist blogs as well as LW2.0 and the EA Forum on RSS. The daily feed is posted in the SSC Discord and on my Wordpress Blog . However the rational feed includes quite alot of articles and many people have requested a more heavily pruned feed. In the past my best-of lists have been well received. I am going to try a monthly best-of list. Lets get to it:

=== The Babble and Prune Sequence by alkjash

This series is probably my favorite concept posted on lesswrong 2.0 so far.

Babble - Babble and Prune model of thought generation: Babble with a weak heuristic to generate many more possibilities than necessary, Prune with a strong heuristic to find a best, or the satisfactory one.

More Babble - Modeling Babble as a random walk on a graph where nodes can represent concepts or words and edges are mental connections. You can improve your babel/prune by increasing connections between different areas or by reducing connections to unproductive clusters. An Analogy between babble/prune perspective on writing and Generative Adversarial Neural Nets.

Prune - People’s filters on their babble are far too strong. The early filters: What you think consciously, what you speak, and what you write down. These filters eliminate all but a trickle of babble. Specific mental techniques to weaken the gates.

NP is the God of Babble. His law is: humans will always be much better at verifying wisdom than producing it. Therefore, go forth and Babble!

=== Politics and Social Dynamics:

An Apology Is A Surrender by Zachary Jacobi – [Recommended largely based on the comments]. If you apologize you should surrender. If you keep fighting its not a real apology. However fake apologies are common. Many commenters point out this implies you often should not surrender to the people demanding an apology. Hence its often rational to ‘fake apologize’.

Nice-manning by The Unit of Caring – How to put yourself in a situation where you are equipped to be nice. Niceness advice: Have a wide audience in mind, When someone is a dick, you can actually just pretend they weren’t, Vary topics a lot.

Whence Comes Nhilism by samzdat – Samzdat wrote a long series of posts on modernity, power and nhilism. References include Nietzche and Seeing Like a State. This peice compresses the series down to 1200 words. However its quite clear and easy to read.

=== Thoughts on Applying Mental Models

Be The Baby With A Hammer by Brian Lui – “Conventional wisdom warns against being fascinated by a new idea and applying it to everything. But the reverse is true; we

should deliberately apply new ideas more than it seems necessary."

Hammers And Nails by alkjash – Hammers are people who apply the same techniques to a variety of problems. Nails are people who apply many techniques to the same few problems. Why its better to either a dedicated nail or a dedicated hammer. Hammers can be underrated, powerful hammers have honed their simple tricks into superweapons.

=== AI Safety

AI: Racing Toward The Brink by Waking Up with Sam Harris – "Sam Harris speaks with Eliezer Yudkowsky about the nature of intelligence, different types of AI, the "alignment problem," IS vs OUGHT, the possibility that future AI might deceive us, the AI arms race, conscious AI, coordination problems, and other topics."

Honest Organizations by Paul Christiano – Goal: Create an honest AI organization. Strategy: The company makes public statements that it promises are not misleading. Employees commit to publicly dispute the statements if they find them misleading. This strategy cannot create trust but it can dramatically amplify initial levels of trust.

=== In Depth Articles: These are recommended but they are extremely long and/or dense

Arbital Postmortem – Arbital was a rationalist startup designed to solve online explanation. Arbital is being shut down, though the site is still online. The founder and main coder of arbital talks about why the site failed, the need to get users more aggressively, attempted pivots and working closely with Eliezer.

Announcement Ai Alignment Prize Winners And Next Round by cousin_it et all – The six winning essays in the AI alignment writing contest. The top prize was given to an essay on Goodhart Taxonomies. "Detailing the possible failures that can arise when optimizing for a proxy instead of the actual goal. Goodhart's Law is simple to understand, impossible to forget once learned, and applies equally to AI alignment and everyday life. While Goodhart's Law is widely known, breaking it down in this new way seems very valuable."

This Review Is Not About Reviewing The Elephant In The Brain by Artir – Very, very detailed review of Robin Hanson's book on hidden motives. Discussion of the core thesis: What exactly is the elephant in the brain? Topics: Signaling Model of Education. Medicine - The key facts to explain and which treatments empirically work. Charity - how people actually donate, the drowning child argument. Diverse thoughts on the underlying theory and which ideas count as confused concepts. [21K words - thats very long even relative to the long articles section]

Rationality Abridged by Quaerendo – "I present to you: Rationality Abridged — a 120-page nearly 50,000-word summary of "Rationality: From AI to Zombies". Yes, it's almost a short book. But it is also true that it's less than 1/10th the length of the original."

Robustness to Scale

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

I want to quickly draw attention to a concept in AI alignment: Robustness to Scale. Briefly, you want your proposal for an AI to be robust (or at least fail gracefully) to changes in its level of capabilities. I discuss three different types of robustness to scale: robustness to scaling up, robustness to scaling down, and robustness to relative scale.

The purpose of this post is to communicate, not to persuade. It may be that we want to bite the bullet of the strongest form of robustness to scale, and build an AGI that is simply not robust to scale, but if we do, we should at least realize that we are doing that.

Robustness to scaling up means that your AI system does not depend on not being too powerful. One way to check for this is to think about what would happen if the thing that the AI is optimizing for were actually maximized. One example of failure of robustness to scaling up is when you expect an AI to accomplish a task in a specific way, but it becomes smart enough to find new creative ways to accomplish the task that you did not think of, and these new creative ways are disastrous. Another example is when you make an AI that is incentivized to do one thing, but you add restrictions that make it so that the best way to accomplish that thing has a side effect that you like. When you scale the AI up, it finds a way around your restrictions.

Robustness to scaling down means that your AI system does not depend on being sufficiently powerful. You can't really make your system still work when it scales down, but you can maybe make sure it fails gracefully. For example, imagine you had a system that was trying to predict humans, and use these predictions to figure out what to do. When scaled up all the way, the predictions of humans are completely accurate, and it will only take actions that the predicted humans would approve of. If you scale down the capabilities, your system may predict the humans incorrectly. These errors may multiply as you stack many predicted humans together, and the system can end up optimizing for some seeming random goal.

Robustness to relative scale means that your AI system does not depend on any subsystems being similarly powerful to each other. This is most easy to see in systems that depend on adversarial subsystems. If part of you AI system is suggest plans, and another part is trying to find problems in those plans, if you scale up the suggester relative to the verifier, the suggester may find plans that are optimized for taking advantage of the verifier's weaknesses.

My current state is that when I hear proposals for AI alignment that do not feel very strongly robust to scale, I become very worried about the plan. Part of this comes from feeling like we are actually very early on a logistic capabilities curve. I thus expect that as we scale up capabilities, we can get eventually get large differences very quickly. Thus, I expect that the scaled up (and partially scaled up) versions to actually happen. However, robustness to scale is very difficult, so it may be that we have to depend on systems that are not very robust, and be careful not to push them too far.

Why we want unbiased learning processes

tl;dr: if an agent has a biased learning process, it may choose actions that are worse (with certainty) for every possible reward function it could be learning.

An agent learns its own reward function if there is a set R of possible reward functions, and there is a learning process P that maps world-histories (and policies) to distributions over R. Thus by interacting with the environment and choosing its own policies, the agent can learn which is the correct reward function it should be maximising.

Given a policy π, a history h, an environment μ,and a reward R, we can compute the expected probability of R:

$$E_{\pi}^{\mu} P(R \mid h).$$

Then a learning process is **unbiased** if that expression is independent of π, and **biased** otherwise. Biased processes are less desirable, as they allow the agent to manipulate the process through its choice of policy.

Simple biased learning process

The most trivial example of a biased learning process is an agent that completely determines its reward by its actions.

Let $R = \{R_0, R_1\}$, let the agent only act once with two actions available, $\{a_0, a_1\}$, (hence a choice of "policy" is a choice of action), and set

$$P(R_0 \mid a_0) = P(R_1 \mid a_1) = 1.$$

Thus the agent can simply choose its reward function through its actions.

Note that some designs are a bit more sophisticated, and don't allow the agent to choose its reward function directly through its actions. But this doesn't matter, if the reward function is a consequence of anything that is a predictable consequence of the agent's actions (eg if the agent can trick/coerce/manipulate a human into saying "yes" or "no", and if P is determined by the human's response, it doesn't matter that P is not defined directly through the agent's actions: it is defined indirectly through them).

[*Note that all* P *that involve learning about external facts are unbiased learning processes, so it's not as if unbiased means trivial*]

Strictly dominated behaviour

Then an agent with a biased learning process that wants to maximise the expectation of the true reward, can sometimes follow strictly dominated policies.

That means that there is are policies π_0 and π_1 , such that for all histories h_i possible given π_i, and all reward function R in R,

$$R(h_1) > R(h_0).$$

And yet the agent will still choose π_0 to maximise reward.

For example, with P and R defined as above, define R_0 and R_1 to be:

	a_0	a_1
R_0	2	3
R_1	0	1

Thus a_1 is always the better action, for both R_0 and for R_1; it is strictly dominant. However, since a_i also determines which reward function is correct, the possible rewards the agent gets are the two bold numbers in the table: 2, by choosing a_0 and hence making R_0 the correct reward function, and 1, by choosing a_1 and hence making R_1 the correct reward function.

Then in order to maximise reward, the agent will choose the strictly dominated policy/action a_0.

Unbiased learning

It's possible to prove that if P is unbiased, then this behaviour won't occur, but doing so involves introducing a bit more definition and machinery that presented here, so I'll defer this to my forthcoming paper.

Note on *expected* dominance

[*Reading the following is not relevant to understanding the main point of this post*]

The dominant policy is defined so that for all $R \in \mathcal{R}$, $R(h_1) > R(h_0)$ for all histories h_i, possible given π_i.

We could instead talk about the expected reward given π_i. But in fact, it makes sense to choose policies which are strictly dominated in the expected reward sense.

For example, let $\pi_\varnothing$ be a policy that does nothing (all rewards stay at 0), and let π_P be the policy that first checks which of R_0 and R_1 is correct (for a given P) and then maximises the correct one. If R_i is maximised, it goes to 1, while the other reward will go to -2.

Assume that the probability of either R_0 or R_1 being correct is 1/2. Then it's clear that π_P is dominated in expectation by $\pi_\varnothing$, since

$$E^{\mu}_{\pi_\varnothing} R_i = 0,$$

$$E^{\mu}_{\pi_P} R_i = 1/2\,(1) + 1/2\,(-2) = -1/2.$$

Yet π_P is clearly the right thing to do, since it allows us to maximise the correct reward (R_i is only negative in worlds where it is not the correct reward).

So an unbiased agent can still choose a policy that is worse for every reward in expectation, if it's confident that the (currently unknown) correct reward will get maximised more by this policy.

Replacing expensive costly signals

I feel like there is a general problem where people signal something using some extremely socially destructive method, and we can conceive of more socially efficient ways to send the same signal, but trying out alternative signals suggests that you might be especially bad at the traditional one. For instance, an employer might reasonably suspect that a job candidate who did a strange online course instead of normal university would have done especially badly at normal university.

Here is a proposed solution. Let X be the traditional signal, Y be the new signal, and Z be the trait(s) being advertised by both. Let people continue doing X, but subsidize Y on top of X for people with very high Z. Soon Y is a signal of higher Z than X is, and understood by the recipients of the signals to be a better indicator. People who can't afford to do both should then prefer Y to X, since Y is is a stronger signal, and since it is more socially efficient it is likely to be less costly for the signal senders.

If Y is intrinsically no better a signal than X (without your artificially subsidizing great Z-possessors to send it) then in the long run Y might only end up as strong a sign as X, but in the process, many should have moved to using Y instead.

(A possible downside is that people may end up just doing both forever.)

For example, if you developed a psychometric and intellectual test that only took half a day and predicted very well how someone would do in an MIT undergraduate degree, you could run it for a while for people who actually do MIT undergraduate degrees, offering prizes for high performance, or just subsidizing taking it at all. After the best MIT graduates say on their CVs for a while that they also did well on this thing and got a prize, it is hopefully an established metric, and an employer would as happily have someone with the degree as with a great result on your test. At which point an impressive and ambitious high school leaver would take the test, modulo e.g. concerns that the test doesn't let you hang out with other MIT undergraduates for four years.

I don't know if this is the kind of problem people actually have with replacing apparently wasteful signaling systems with better things. Or if this doesn't actually work after thinking about it for more than an hour. But just in case.

Crypto autopsy reply

(X-posted from my FB post:
https://www.facebook.com/alexei.andreev.3/posts/1403550339754401)

Reply to Eliezer's post on crypto:
https://www.facebook.com/yudkowsky/posts/10156147605134228
Which itself is a response to Scott Alexander's post:
https://www.lesserwrong.com/…/Ma…/a-lesswrong-crypto-autopsy

One thing I haven't seen discussed in either of the posts is the social aspect. When Bitcoin was created, LW was a hub for a lot of pretty smart contrarians. Just with that info alone, I'd give any crazy-seeming idea that surfaced there and wasn't immediately debunked pretty good odds of being vaguely correct and likely ahead of its time. (I think there is a good chance LW 2.0 might become that again.)

So I agree with Scott, the tragedy is that our very smart peers brought this idea to the community, and the community vaguely agreed that it's a good idea, but failed to execute. (Just like our community mostly agrees that cryo is a good idea, but also gave birth to the term cryostination.)

Eliezer's counter-claim about efficient markets seems weak to me. It was clear at the time that this was a very niche idea, and that not a lot of people have looked into it. It was also a difficult idea to digest: it required a solid understanding of the economics and the technical parts. Many people who care about making money are not technical, and can't evaluate cutting-edge technical ideas. I think Bitcoin was very clearly out of reach of efficient markets. (By Eliezer's reasoning, AGI would be a bad idea around 2010, and so would be cryo even now. Because if they were as good as they seemed, surely people who were interested in power/money would be putting way more effort into AGI, and people who were interested in not dying would be doing cryo right now.)

I don't think I saw those early Bitcoin posts on LW. My own story was that my brother asked me around 2012: "Hey, have you heard about this Bitcoin thing? I just read about it on Reddit. Seems cool, we should buy some!" And I laughed and said it was the dumbest thing ever.

Less than a year later he brought it up again, and it was clear by that time that I made a mistake. That's when my rationality training kicked in. "Ouch, I though. He was totally right, and I completely dismissed this idea. I should not make the same mistake again!" By that time my social network was talking about it too. Getting a consistent signal like that from multiple sources to me means it's something worth paying serious attention to. So, I bought some Bitcoins back in 2013. Of course, I held it in Mt. Gox, so that did me no good. After I lost it all, I bought more on Cryptsy, which met the same fate.

Then came Ethereum. By that time I had sufficient interest in crypto to take the time to read and understand what it was about, and how it was different. (This was before there were 100s of ICOs each day, but there were still a good number of new coins being launched.) But Ethereum had a clear strong social signal! So while I didn't totally trust myself to evaluate this opportunity solely on my own, I listened closely to what my network was saying about it. Most people thought it was a great buy and their reasoning made sense to me. In my mind this was very similar to the original Bitcoin scenario. And this time I didn't make the same mistake. (I wonder what other

rationalists did regarding Ethereum. Surely most have heard about it. What prevented them from buying it? Did they not learn their lesson with Bitcoin?)

Then came Tezos less than a year ago. And, again, my social network was talking about. And, again, I looked into it and it seemed solid. Even though by that time there were a *lot* of ICOs going on, this one had a very clear signal through my social network. So I invested even more than I did in Ethereum's ICO. Too early to say for sure how this will turn out until it's actually launched, but so far it's a 10x return. (And I actually posted about this ICO: https://www.facebook.com/alexei.andreev.3/posts/1215850181857752)

My point is: if you're a rationalist, you've surrounded yourself with some *very smart* people. Listen to them. In aggregate, take their ideas seriously even when they might not take their own ideas seriously. Like Eliezer wrote in his recent sequence: in any given field, it's much easier to find the people with the correct contrarian opinion than it is to develop a correct contrarian opinion on your own. There is no reason why that couldn't have (or still can't) apply to the crypto space as well.

Inconvenience Is Qualitatively Bad

My most complicated cookie recipe has four layers. Two of these require stovetop cooking, and the other two require the use of the oven separately before the nearly-complete cookies are baked in yet a third oven use, for a total of three different oven temperatures. I have to separate eggs. I have to remember to put butter out hours in advance so it'll be softened when I get underway. Spreading the fruit neatly and then the almond goop on top of that without muddling the layers is finicky and almost none of the steps parallelize well.

They're delicious, but at what cost?

People who don't cook as a hobby would never, ever make these cookies. And this is reasonable. They shouldn't. On most days I shouldn't either. They are staggeringly inconvenient.

But they're made up of individual steps that you could mostly figure out if you really wanted to. Lots and lots of little steps. This is why I want to scream whenever I hear someone try to *add steps* to someone else's life. Especially if they say "just".

"Just" Google it. "Just" rinse out your recyclables. "Just" add another thing to remember and another transition to your to-do list and another obligation to feel guilt about neglecting and another source of friction between you and your real priorities. It "just" takes a minute. Don't you *care*?

Anyone who didn't have any immune defense against things that just take a minute would spend fifteen hours a day on one-minute tasks, if expending the energy required to switch between the tasks didn't kill them before it got that bad. But "it would be inconvenient" doesn't tend to feel like a solid rebuttal - to either party; the one attempting to impose can just reiterate "but it'll only take a *minute*".

Everyone needs algorithms to cut down on inconveniences.

Some I am aware of:

- Chunking. Things feel less inconvenient (and accordingly are) if they are understood in batches, as one thing and not thirty. (This is related, I think, to a lot of manifestations of executive dysfunction - chunking doesn't work as well.) People naturally do more of this with things they're good at - I think a lot of being good at things *just is* being able to take them in larger chunks and finding larger amounts of the thing trivial.
- Delegating. For some people delegating is itself inconvenient but if you delegate enough things it can be useful on balance. Often costly in other ways too - quality, customization, money.
- Straight-up rejecting impositions, in whole ("I just won't recycle") or in part ("I'll recycle but no way am I washing out bean cans"). Pick what to reject at whim, from certain sources, or by another discrimination mechanism. Rejecting impositions from interactive humans as opposed to generic announcements or oneself requires social grace or a willingness to do without it.

Hufflepuff Cynicism on Crocker's Rule

Yesterday, I mainly talked about Hufflepuff Cynicism from the cynic's end. However, there's a lot to be said about the receiving end. Hufflepuff cynicism can come off as a very patronizing strategy. Is this a point against it?

In the original conversation where I came up with the idea of Hufflepuff cynicism, I was talking about norms for aspiring rationalists around trying to get other people to be more rational. Maybe we agree that double crux is a good conversation procedure, but should we *try to convince* someone of that? Should we try to get them to double-crux with us about it? Maybe we believe you should bet or update when a disagreement hasn't been resolved, but what should we do with a disagreement about the bet-or-update rule?

My argument from the Hufflepuff Cynicism side was in favor of chesterton-fencing such disagreements. Don't try to convince others about rationality norms; at least, stop after the first explanation falls on deaf ears. Instead, figure out why the person isn't already following the norm. It seems likely that there's some important reason; if you can figure it out, maybe you can come up with a better norm which would address the concern (in much the same way bet-or-update addresses objections to the simpler strategies "bet on disagreements" or "talk out disagreements until you converge").

To my surprise, not everyone wants to be treated so carefully. Some people find this attitude patronizing or overly cautious, and request that I *just tell them what they are doing wrong*, possibly telling them *more than just one time* if they don't get it the first time. This is, more or less, an invocation of Crocker's Rule.

To quote the sl4 wiki on Crocker's Rules:

> Declaring yourself to be operating by "Crocker's Rules" means that other people are allowed to optimize their messages for information, not for being nice to you. Crocker's Rules means that you have accepted full responsibility for the operation of your own mind - if you're offended, it's your fault. Anyone is allowed to call you a moron and claim to be doing you a favor. (Which, in point of fact, they would be. One of the big problems with this culture is that everyone's afraid to tell you you're wrong, or they think they have to dance around it.) Two people using Crocker's Rules should be able to communicate all relevant information in the minimum amount of time, without paraphrasing or social formatting. Obviously, don't declare yourself to be operating by Crocker's Rules unless you have that kind of mental discipline.
>
> **Note that Crocker's Rules does *not* mean you can insult people; it means that *other* people don't have to worry about whether *they* are insulting *you*.** Crocker's Rules are a discipline, not a privilege. Furthermore, taking advantage of Crocker's Rules does not imply reciprocity. How could it? Crocker's Rules are something you do for yourself, to maximize information received - *not* something you grit your teeth over and do as a favor.

(This seems like really just one rule to me, so I tend to call it Crocker's Rule.)

A problem I've encountered, which has reinforced my Hufflepuff Cynicism, is that people can invoke Crocker's Rule and then get upset about feedback I give anyway.

My advice is this: Crocker's Rule is a promise not to punish others for giving you negative feedback. If you don't want to be patronized by Hufflepuff Cynics like myself, don't make that promise unless you're sure you can keep it. Instead, *show others through your words and actions over an extended period of time* that you are both able and happy to accept negative feedback. Don't make a standing request for negative feedback. Ask for it explicitly again and again, and thank others for giving it to you. If you can't thank them genuinely, you've learned something about yourself and can adapt accordingly.

But, maybe I'm too much of a Hufflepuff Cynic. I don't know. Maybe I should... hold people to their own standards, sometimes...?

Mapping the Archipelago

I got excited reading Meta-tations on Moderation: Towards Public Archipelago for two reasons: there's a clear island of the archipelago I've been mostly avoiding on LessWrong, and the whole place has been growing at a high enough rate to demand fracturing.

Since we have the chance to direct the growth of the brand new archipelago, let's start a discussion down one level of meta: what specific islands do you want to see? Second, how should discussion and moderation norms differ between them?

Three islands of current LW, according to me:

1. AI Risk: Serious discussion for serious folk. No smiles allowed.
2. Instrumental Rationality: The means to get to the ends. Whatever they might be.
3. Fluff and Fiction: Blood for the Art God! Fun over fact.

(Guess which island I avoid.)

Arguments about fast takeoff

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://sideways-view.com/2018/02/24/takeoff-speeds/

I expect "slow takeoff," which we could operationalize as the economy doubling over some 4 year interval before it doubles over any 1 year interval. Lots of people in the AI safety community have strongly opposing views, and it seems like a really important and intriguing disagreement. I feel like I don't really understand the fast takeoff view.

(Below is a short post copied from Facebook. The link contains a more substantive discussion. See also: AI impacts on the same topic.)

I believe that the disagreement is mostly about what happens *before* we build powerful AGI. I think that weaker AI systems will already have radically transformed the world, while I believe fast takeoff proponents think there are factors that makes weak AI systems radically less useful. This is strategically relevant because I'm imagining AGI strategies playing out in a world where everything is already going crazy, while other people are imagining AGI strategies playing out in a world that looks kind of like 2018 except that someone is about to get a decisive strategic advantage.

Here is my current take on the state of the argument:

The basic case for slow takeoff is: "it's easier to build a crappier version of something" + "a crappier AGI would have almost as big an impact." This basic argument seems to have a great historical track record, with nuclear weapons the biggest exception.

On the other side there are a bunch of arguments for fast takeoff, explaining why the case for slow takeoff doesn't work. If those arguments were anywhere near as strong as the arguments for "nukes will be discontinuous" I'd be pretty persuaded, but I don't yet find any of them convincing.

I think the best argument is the historical analogy to humans vs. chimps. If the "crappier AGI" was like a chimp, then it wouldn't be very useful and we'd probably see a fast takeoff. I think this is a weak analogy, because the discontinuous progress during evolution occurred on a metric that evolution wasn't really optimizing: groups of humans can radically outcompete groups of chimps, but (a) that's almost a flukey side-effect of the individual benefits that evolution is actually selecting on, (b) because evolution optimizes myopically, it doesn't bother to optimize chimps for things like "ability to make scientific progress" even if in fact that would ultimately improve chimp fitness. When we build AGI we will be optimizing the chimp-equivalent-AI for usefulness, and it will look nothing like an actual chimp (in fact it would almost certainly be enough to get a decisive strategic advantage if introduced to the world of 2018).

In the linked post I discuss a bunch of other arguments: people won't be trying to build AGI (I don't believe it), AGI depends on some secret sauce (why?), AGI will improve radically after crossing some universality threshold (I think we'll cross it way before AGI is transformative), understanding is inherently discontinuous (why?), AGI will be much faster to deploy than AI (but a crappier AGI will have an intermediate

deployment time), AGI will recursively improve itself (but the crappier AGI will recursively improve itself more slowly), and scaling up a trained model will introduce a discontinuity (but before that someone will train a crappier model).

I think that I don't yet understand the core arguments/intuitions for fast takeoff, and in particular I suspect that they aren't on my list or aren't articulated correctly. I am very interested in getting a clearer understanding of the arguments or intuitions in favor of fast takeoff, and of where the relevant intuitions come from / why we should trust them.

Write a Thousand Roads to Rome

Epistemological Status: Pretty sure I'm on to something here, also very sure I'm restating the obvious, utterly confident that restating the obvious is the point.

Sometimes a piece of writing gets two very different responses. Half the commenters say something like "this is really obvious and a waste of time to write" and half of them say something like "this is revolutionary and an amazing insight I never would have reached on my own, thank you!" I think these posts that get this reaction are vital, possibly even more useful in some ways than a post breaking new ground.

First, a preamble describing how my brain is finicky.

When I was first learning Calculus in college, I had a hard time conceptually understanding what a derivative was. I could do the operation most of the time, but didn't know what that operation represented or what real thing those symbols on the page described. Someone suggested it was the rate of change and the second derivative was the rate of change in the rate of change, but that didn't mean anything to me. Someone else suggested it was velocity and the second derivative was acceleration, but I was honestly not great at physics either. I think someone suggested something about stock prices or economics, but I didn't have enough econ knowledge at the time to remember what exactly they said.

Then my professor recognized one of the equations I'd been messing around with on some scrap paper as describing the attack bonus from a game, and pointed out that the derivative of that would be how much better the character would get each level and the second derivative would be useful if characters got stronger at different rates.

"Oh, like an onion knight from Final Fantasy?" I asked.

The professor just gave a confused expression, but the concept had already clicked for me and later that week I was using derivatives at my gaming sessions as well as suddenly doing much better in my physics classes. The moral of the story is that if you want me to understand something, the best way is usually to relate it to a roleplaying game. I'm not the only one who thinks like this- a close friend of mine had no idea why I was worried about intelligence explosions until I reminded him of the Intellect Potion exploit in Morrowind. Yes, this is a silly way for a brain to work, but I think we all agree brains don’t necessarily work in sensible ways.

Next, a suggestion that most brains are finicky.

SlateStarCodex’s What Universal Human Experiences Are You Missing Without Realizing It? is a gift that keeps on giving. The first example is that you can be completely unable to smell, and not realize this until asked to write about how a peach smells. What makes that relevant here is that they must have been asked about smells before and they offered up an answer copied from people around them, but the particular way that question got asked this time caught them short and they realized they couldn't smell.

The idea of learning styles is a fad that came and went in a flash, but the useful parts struck a chord with lots of people. My brother remembers things best if he listens to them and repeats them aloud. I remember things best if I learn them as part of a story and then write them down on an index card. My roommate remembers things best if

they draw a picture or diagram of some kind. These kinds of mental differences are really common, and there is no single best format for information to be presented in.

A written article revealing revolutionary insights will never get absorbed by my brother until the audiobook comes out. If I want him to know a thing, then writing better articles with better insights isn't actually what I need to do; I need to read the ones I have aloud.

Are we raising the sanity waterline, or are we building a sanity waterspout?

If a few people in your society are literate, then they can copy the holy books to pass down knowledge for future generations and maybe even do a little written debate with each other in the margins. If more people in your society are literate, then you can start using a printing press for newspapers and leaflets. If most people in your society are literate, you get wikipedia. Some skills have incredible compounding effects when more people in a society have them; pick your favourite three rationalist techniques, and imagine a society where those skills are as common as literacy is in America today. I suspect that an alternate universe where every adult knew that arguments aren't soldiers, that beliefs should pay rent, and that nobody is perfect but everything is commensurable would be doing really well compared to us, even if it cost them some new breakthroughs!

Actually, I don't think writing these restatements costs you that many new concepts. Writing out an explanation of an idea one already holds is easier for many people than coming up with a new idea, and I find the process of breaking down a notion to explain it to someone else often improves my own understanding of the original concept. Small changes or additions to an idea can result in something new; Nobody is Perfect, Everything is Commensurable is at least half derived from Money: the Unit of Caring, but the other half seems to have done good work by providing contrast. I've never had a math professor who had made an original discovery in mathmatics but many of them taught me useful things, and Khan Academy isn't going to be writing new proofs anytime soon but they're still doing valuable work by providing another way for people to learn the things that already are known. There isn't a clean tradeoff between restatements and new concepts because these use different skills.

I'm not suggesting that repeating things using different words is always better than describing a new insight; I'd rather have one Eliezer Yudkowsky working on the big problems than a hundred slightly more rational Screwtapes, because sometimes what matters is the highest intellectual peak your society can reach even if it's just one person. For a lot of problems today what matters is the peak and you want to boost one or two people to those lofty heights. (I think of these as "sanity waterspout" problems.) That said, if I want to learn multivariate calculus or basic game theory there will be at least a dozen different ways of learning it, from a video with pictures to a terse few pages of a textbook to some goofy edutainment videogame. This is a good thing, and I wish rationality was more like this. Think of your favourite technique: how many different formats can it be found in? Alternately, think of your favourite format for learning something: how many techniques and concepts can be found in it?

For every concept we want more people to understand, we should want it explained in more ways. If we want everyone to be a little more rational, then we need to put concepts into mediums that everyone can understand.

Toward a New Technical Explanation of Technical Explanation

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

A New Framework

(Thanks to Valentine for a discussion leading to this post, and thanks to CFAR for running the CFAR-MIRI cross-fertilization workshop. Val provided feedback on a version of this post. Warning: fairly long.)

Eliezer's *A Technical Explanation of Technical Explanation*, and moreover the sequences as a whole, used the best technical understanding of practical epistemology available at the time* -- the Bayesian account -- to address the question of how humans can try to arrive at better beliefs in practice. The sequences also pointed out several holes in this understanding, mainly having to do with logical uncertainty and reflective consistency.

MIRI's research program has since then made major progress on logical uncertainty. The new understanding of epistemology -- the theory of logical induction -- generalizes the Bayesian account by eliminating the assumption of logical omniscience. Bayesian belief updates are recovered as a special case, but the dynamics of belief change are non-Bayesian in general. While it might not turn out to be the last word on the problem of logical uncertainty, it has a large number of desirable properties, and solves many problems in a unified and relatively clean framework.

It seems worth asking what consequences this theory has for practical rationality. Can we say new things about what good reasoning looks like in humans, and how to avoid pitfalls of reasoning?

First, I'll give a shallow overview of logical induction and possible implications for practical epistemic rationality. Then, I'll focus on the particular question of *A Technical Explanation of Technical Explanation* (which I'll abbreviate TEOTE from now on). Put in CFAR terminology, I'm seeking a gears-level understanding of gears-level understanding. I focus on the intuitions, with only a minimal account of how logical induction helps make that picture work.

Logical Induction

There are a number of difficulties in applying Bayesian uncertainty to logic. No computable probability distribution can give non-zero measure to the logical tautologies, since you can't bound the amount of time you need to think to check whether something is a tautology, so updating on provable sentences always means updating on a set of measure zero. This leads to convergence problems, although there's been recent progress on that front.

Put another way: Logical consequence is deterministic, but due to Gödel's first incompleteness theorem, it is like a stochastic variable in that there is no computable procedure which correctly decides whether something is a logical consequence. This means that any computable probability distribution has infinite Bayes loss on the question of logical consequence. Yet, because the question is actually deterministic, we know how to point in the direction of better distributions by doing more and more consistency checking. This puts us in a puzzling situation where we want to improve the Bayesian probability distribution by doing a kind of non-Bayesian update. This was the two-update problem.

You can think of logical induction as supporting a set of hypotheses which are about ways to shift beliefs as you think longer, rather than fixed probability distributions which can only shift in response to evidence.

This introduces a new problem: how can you score a hypothesis if it keeps shifting around its beliefs? As TEOTE emphasises, Bayesians outlaw this kind of belief shift for a reason: requiring predictions to be made in advance eliminates hindsight bias. (More on this later.) So long as you understand exactly what a hypothesis predicts and what it does not predict, you can evaluate its Bayes score and its prior complexity penalty and rank it objectively. How do you do this if you don't know all the consequences of a belief, and the belief itself makes shifting claims about what those consequences are?

The logical-induction solution is: set up a prediction market. A hypothesis only gets credit for contributing to collective knowledge by moving the market in the right direction early. If the market's odds on prime numbers are currently worse than those which the prime number theorem can provide, a hypothesis can make money by making bets in that direction. If the market has already converged to those beliefs, though, a hypothesis can't make any more money by expressing such beliefs -- so it doesn't get any credit for doing so. If the market has moved on to even more accurate rules of thumb, a trader would only lose money by moving beliefs back in the direction of the prime number theorem.

Mathematical Understanding

This provides a framework in which we can make sense of mathematical labor. For example, a common occurrence in combinatorics is that there is a sequence which we can calculate, such as the catalan numbers, by directly counting the number of objects of some specific type. This sequence is boggled at like data in a scientific experiment. Different patterns in the sequence are observed, and hypotheses for the continuation of these patterns are proposed and tested. Often, a significant goal is the construction of a closed form expression for the sequence.

This looks just like Bayesian empiricism, except for the fact that *we already have a hypothesis which entirely explains the observations.* The sequence is constructed from a definition which mathematicians made up, and which thus assigns 100% probability to the observed data. What's going on? It is possible to partially explain this kind of thing in a Bayesian framework by acting *as if* the true formula were unknown and we were trying to guess where the sequence came from, but this doesn't explain everything, such as why finding a closed form expression would be important.

Logical induction explains this by pointing out how different time-scales are involved. Even if all elements of the sequence are calculable, a new hypothesis can get credit

for calculating them faster than the brute-force method. Anything which allows one to produce correct answers faster contributes to the efficiency of the prediction market inside the logical inductor, and thus, to the overall mathematical understanding of a subject. This cleans up the issue nicely.

What other epistemic phenomena can we now understand better?

Lessons for Aspiring Rationalists

Many of these could benefit from a whole post of their own, but here's some fast-and-loose corrections to Bayesian epistemology which may be useful:

- Hypotheses need not make predictions about everything. Because hypotheses are about how to *adjust* your odds as you think longer, they can leave most sentences alone and focus on a narrow domain of expertise. Everyone was already doing this in practice, but the math of Bayesian probability theory requires each hypothesis to make a prediction about every observation, if you actually look at it. Allowing a hypothesis to remain silent on some issues in standard Bayesianism can cause problems: if you're not careful, a hypothesis can avoid falsification by remaining silent, so you end up incentivising hypotheses to remain mostly silent (and you fail to learn as a result). Prediction markets are one way to solve this problem.
- Hypotheses buy and sell at the *current price*, so they take a hit for leaving a now-unpopular position which they initially supported (but less of a hit than if they'd stuck with it) or coming in late to a position of growing popularity. Other stock-market type dynamics can occur.
- Hypotheses can be like object-level beliefs or meta-level beliefs: you can have a hypothesis about how you're overconfident, which gets credit for smoothing your probabilities (if this improves things on average). This allows you to take into account beliefs about your calibration without getting *too* confused about Hofstadter's-law type paradoxes.

You may want to be a bit careful and Chesterton-fence existing Bayescraft, though, because some things are still better about the Bayesian setting. I mentioned earlier that Bayesians don't have to worry so much about hindsight bias. This is closely related to the problem of old evidence.

Old Evidence

Suppose a new scientific hypothesis, such as general relativity, explains a well-know observation such as the perihelion precession of mercury better than any existing theory. Intuitively, this is a point in favor of the new theory. However, the probability for the well-known observation was already at 100%. How can a previously-known statement provide new support for the hypothesis, as if we are re-updating on evidence we've already updated on long ago? This is known as the problem of old evidence, and is usually levelled as a charge against Bayesian epistemology. However, in some sense, the situation is worse for logical induction.

A Bayesian who endorses Solomonoff induction can tell the following story: Solomonoff induction is the right theory of epistemology, but we can only approximate it, because it is uncomputable. We approximate it by searching for hypotheses, and computing their posterior probability retroactively when we find new ones. It only makes sense

that when we find a new hypothesis, we calculate its posterior probability by multiplying its prior probability (based on its description length) by the probability it assigns to all evidence so far. That's Bayes' Law! The fact that we already knew the evidence is not relevant, since our approximation didn't previously include this hypothesis.

Logical induction speaks against this way of thinking. The hypothetical Solomonoff induction advocate is assuming one way of approximating Bayesian reasoning via finite computing power. Logical induction can be thought of as a different (more rigorous) story about how to approximate intractible mathematical structures. In this new way, *propositions are bought or sold at market prices at the time.* If a new hypothesis is discovered, it can't be given any credit for 'predicting' old information. The price of known evidence is already at maximum -- you can't gain any money by investing in it.

There are good reasons to ignore old evidence, especially if the old evidence has biased your search for new hypotheses. Nonetheless, it doesn't seem right to *totally* rule out this sort of update.

I'm still a bit puzzled by this, but I think the situation is improved by understanding gears-level reasoning. So, let's move on to the discussion of TEOTE.

Gears of Gears

As Valentine noted in his article, it is somewhat frustrating how the overall idea of gears-level understanding seems so clear while remaining only heuristic in definition. It's a sign of a ripe philosophical puzzle. If you don't feel you have a good intuitive grasp of what I mean by "gears level understanding", I suggest reading his post.

Valentine gives three tests which point in the direction of the right concept:

1. Does the model pay rent? If it does, and if it were falsified, how much (and how precisely) could you infer other things from the falsification?
2. How incoherent is it to imagine that the model is accurate but that a given variable could be different?
3. If you knew the model were accurate but you were to forget the value of one variable, could you rederive it?

I already named one near-synonym for "gears", namely "technical explanation". Two more are "inside view" and Elon Musk's notion of reasoning from first principles. The implication is supposed to be that gears-level understanding is in some sense better than other sorts of knowledge, but this is decidedly not supposed to be valued to the exclusion of other sorts of knowledge. Inside-view reasoning is traditionally supposed to be combined with outside-view reasoning (although Elon Musk calls it "reasoning by analogy" and considers it inferior, and much of Eliezer's recent writing warns of its dangers as well, while allowing for its application to special cases). I suggested the terms gears-level & policy-level in a previous post (which I actually wrote after most of this one).

Although TEOTE gets close to answering Valentine's question, it doesn't quite hit the mark. The definition of "technical explanation" provided there is a theory which strongly concentrates the probability mass on specific predictions and rules out others. It's clear that a model can do this without being "gears". For example, my

model might be that whatever prediction the Great Master makes will come true. The Great Master can make very detailed predictions, but I don't know how they're generated. I lack the understanding associated with the predictive power. I might have a strong outside-view reason to trust the Great Master: their track record on predictions is immaculate, their Bayes-loss miniscule, their calibration supreme. Yet, I lack an inside-view account. I can't derive their predictions from first principles.

Here, I'm siding with David Deutsch's account in the first chapter of *The Fabric of Reality*. He argues that understanding and predictive capability are distinct, and that understanding is about having good explanations. I may not accept his whole critique of Bayesianism, but that much of his view seems right to me. Unfortunately, he doesn't give a *technical* account of what "explanation" and "understanding" could be.

First Attempt: Deterministic Predictions

TEOTE spends a good chunk of time on the issue of making predictions in advance. According to TEOTE, this is a human solution to a human problem: you make predictions in advance so that you can't make up what predictions you could have made after the fact. This counters hindsight bias. An ideal Bayesian reasoner, on the other hand, would never be tempted into hindsight bias in the first place, and is free to evaluate hypotheses on old evidence (as already discussed).

So, is gears-level reasoning just pure Bayesian reasoning, in which hypotheses have strictly defined probabilities which don't depend on anything else? Is outside-view reasoning the thing logical induction adds, by allowing the beliefs of a hypothesis to shift over time and to depend on on the wider market state?

This isn't quite right. An ideal Bayesian can still learn to trust the Great Master, based on the reliability of the Great Master's predictions. Unlike a human (and unlike a logical inductor), the Bayesian will at all times have in mind all the possible ways the Great Master's predictions *could* have become so accurate. This is because a Bayesian hypothesis contains a full joint distribution on all events, and an ideal Bayesian reasons about all hypotheses at all times. In this sense, the Bayesian always operates from an inside view -- it cannot trust the Great Master without a hypothesis which correlates the Great Master with the world.

However, it is possible that this correlation is introduced in a very simple way, by ruling out cases where the Great Master and reality disagree without providing any mechanism explaining how this is the case. This may have low prior probability, but gain prominence due to the hit in Bayes-score other hypotheses are taking for not taking advantage of this correlation. It's not a bad outcome given the epistemic situation, but it's not gears-level reasoning, either. So, being fully Bayesian or not isn't *exactly* what distinguishes whether advanced predictions are needed. What is it?

I suggest it's this: *whether the hypothesis is well-defined, such that anyone can say what predictions it makes without extra information.* In his post on gears, Valentine mentions the importance of "*how deterministically interconnected the variables of the model are*". I'm pointing at something close, but importantly distinct: how deterministic the *predictions* are. You know that a coin is very close to equally likely to land on heads or tails, and from this you can (if you know a little combinatorics) compute things like the probability of getting exactly three heads if you flip the coin five times. Anyone with the same knowledge would compute the same thing. The

model includes probabilities inside it, but how those probabilities flow is perfectly deterministic.

This is a notion of objectivity: a wide variety of people can agree on what probability the model assigns, despite otherwise varied background knowledge.

If a model is well-defined in this way, it is very easy (Bayesian or no) to avoid hindsight bias. You cannot argue about how you could have predicted some result. Anyone can sit down and calculate.

The hypothesis that the Great Master is always correct, on the other hand, does not have this property. Nobody but the Great Master can say what that hypothesis predicts. If I know what the Great Master says about a particular thing, I can evaluate the accuracy of the hypothesis; but, this is special knowledge which I need in order to give the probabilities.

The Bayesian hypothesis which simply forces statements of the Great Master to correlate with the world is somewhat more gears-y, in that there's a probability distribution which can be written down. However, this probability distribution is a complicated mish-mosh of the Bayesian's other hypotheses. So, predicting what it would say requires extensive knowledge of the private beliefs of the Bayesian agent involved. This is typical of the category of non-gears-y models.

Objection: Doctrines

Infortunately, this account doesn't totally satisfy what Valentine wants.

Suppose that, rather than making announcements on the fly, the Great Master has published a set of fixed Doctrines which his adherents memorize. As in the previous thought experiment, the word of the Great Master is infallible; the application of the Doctrines always leads to correct predictions. However, the contents of the Doctrines appears to be a large mish-mosh of rules with no unifying theme. Despite their apparent correctness, they fail to provide any understanding. It is as if a physicist took all the equations in a physics text, transformed them into tables of numbers, and then transported those tables to the middle ages with explanations of how to use the tables (but none of where they come from). Though the tables work, they are opaque; there is no insight as to how they were determined.

The Doctrines are a deterministic tool for making predictions. Yet, they do not seem to be a gears-level model. Going back to Valentine's three tests, the Doctrines fail test three: we could erase any one of the Doctrines and we'd be unable to rederive it by how it fit together with the rest. Hence, the Doctrines have almost as much of a "trust the Great Master" quality as listening to the Great Master directly -- the disciples would not be able to derive the Doctrines for themselves.

Second Attempt: Proofs, Axioms, & Two Levels of Gears

My next proposal is that *having a gears-level model is like knowing the proof*. You might believe a mathematical statement because you saw it in a textbook, or because

you have a strong mathematical intuition which says it must be true. But, you don't have the gears until you can prove it.

This subsumes the "deterministic predictions" picture: a model is an axiomatic system. If we know all the axioms, then we can in theory produce all the predictions ourselves. (Thinking of it this way introduces a new possibility, that the model may be well-defined but we may be unable to *find* the proofs, due to our own limitations.) On the other hand, we don't have access to the axioms of the theory embodied by the Great Master, and so we have no hope of seeing the proofs; we can only observe that the Great Master is always right.

How does this help with the example of the Doctrines?

The concept of "axioms" is somewhat slippery. There are many equivalent ways of axiomatizing any given theory. We can often flip views between what's taken as an axiom vs what's proved as a theorem. However, the most elegant set of axioms tends to be preferred.

So, we *can* regard the Doctrines as one long set of axioms. If we look at them that way, then adherents of the Great Master have a gears-level understanding of the Doctrines if they can successfully apply them as instructed.

However, the Doctrines are not an elegant set of axioms. So, viewing them in this way is very unnatural. It is more natural to see them as a set of assertions which the Great Master has produced by some axioms unknown to us. In this respect, we "can't see the proofs".

In the same way, we can consider flipping any model between the axiom view and the theorem view. Regarding the model as axiomatic, to determine whether it is gears-level we only ask whether its predictions are well-defined. Regarding in in "theorem view", we ask if we know how *the model itself* was derived.

Hence, two of Valentine's desirable properties of a gears-level model can be understood as the same property applied at different levels:

- *Determinism*, which is Val's property #2, follows from requiring that we can see the derivations within the model.
- *Reconstructability,* Val's property #3, follows from requiring that we can see the derivation *of* the model.

We might call the first level of gears "made out of gears", and the second level "made *by* gears" -- the model itself being constructed via a known mechanism.

If we change our view so that a scientific theory is a "theorem", what are the "axioms"? Well, there are many criteria which are applied to scientific theories in different domains. These criteria could be thought of as pre-theories or meta-theories. They encode the hard-won wisdom of a field of study, telling us what theories are likely to work or fail in that field. But, a very basic axiom is: we want a theory to be *the simplest theory consistent with all observations.* The Great Master's Doctrines cannot possibly survive this test.

To give a less silly example: if we train up a big neural network to solve a machine learning problem, the predictions made by the model are deterministic, predictable from the network weights. However, someone else who knew all the principles by which the network was created would nonetheless train up a very different neural

network -- unless they use the very same gradient descent algorithm, data, initial weights, and number and size of layers.

Even if they're the same in all those details, and so reconstruct the same neural network *exactly,* there's a significant sense in which they can't *see* how the conclusion follows inevitably from the initial conditions. It's less doctrine-y than being handed a neural network, but it's more doctrine-y than understanding the structure of the problem and why almost any neural network achieving good performance on the task will have certain structures. Remember what I said about mathematical understanding. There's always another level of "being able to see why" you can ask for. Being able to reproduce the proof is different from being able to explain why the proof has to be the way it is.

Exact Statement?

Gears-y ness is a matter of degree, and there are several interconnected things we can point at, and a slippage of levels of analysis which makes everything quite complicated.

In the ontology of math/logic, we can point at whether you can see the proof of a theorem. There are several slippages which make this fuzzier than it may seem. First: do you derive it only form the axioms, or do you use commonly known theorems and equivalences (which you may or may not be able to prove if put on the spot)? There's a long continuum between what one mathematician might say to another as proof and a formal derivation in logic. Second: how well can you see why the proof has to be? This is the spectrum between following each proof step individually (but seeing them as almost a random walk) vs seeing the proof as an elementary application of a well-known technique. Third: we can start slipping the axioms. There are small changes to the axioms, in which one thing goes from being an axiom to a theorem and another thing makes the opposite transition. There are also large changes, like formalizing number theory via the Peano axioms vs formalizing it in set theory, where the entire description language changes. You need to translate from statements of number theory to statements of set theory. Also, there is a natural ambiguity between taking something as an axiom vs requiring it as a condition in a theorem.

In the ontology of computation, we can point at knowing the output of a machine vs being able to run it by hand to show the output. This is a little less flexible than the concept of mathematical proof, but essentially the same distinction. Changing the axioms is like translating the same algorithm to a different computational formalism, like going between Turing machines and lambda calculus. Also, there is a natural ambiguity between a program vs an input: when you run program XYZ with input ABC on a universal Turing machine, you input XYZABC to the universal turing machine; but, you can also think of this as running program XY on input ZABC, or XYZA on input BC, et cetera.

In the ontology of ontology, we could say "can you see why this has to be, from the structure of the ontology describing things?" "Ontology" is less precise than the previous two concepts, but it's clearly the same idea. A different ontology doesn't necessarily support the same conclusions, just like different axioms don't necessarily give the same theorems. However, the reductionist paradigm holds that the ontologies we use should all be consistent with one another (under some translation between the ontologies). At least, aspire to be eventually consistent. Analogous to axiom/assumption ambiguity and program/input ambiguity, there is ambiguity

between an ontology and the cognitive structure which created and justifies the ontology. We can also distinguish more levels; maybe we would say that an ontology doesn't make predictions directly, but provides a language for stating models, which make predictions. Even longer chains can make sense, but it's all subjective divisions. However, unlike the situation in logic and computation, we can't expect to articulate the full support structure for an ontology; it is, after all, a big mess of evolved neural mechanisms which we don't have direct access to.

Having established that we can talk about the same things in all three settings, I'll restrict myself to talking about ontologies.

Two-level definition of gears: A conclusion is gears-like with respect to a particular ontology to the extent that you can "see the derivation" in that ontology. A conclusion is gears-like without qualification to the extent that you can also "see the derivation" of the ontology itself. This is contiguous with gears-ness relative to an ontology, because of the natural ambiguity between programs and their inputs, or between axioms and assumptions. For a given example, though, it's generally more intuitive to deal with the two levels separately.

Seeing the derivation: There are several things to point at by this phrase.

- As in TEOTE, we might consider it important that a model make *precise* predictions. This could be seen as a prerequisite of "seeing the derivation": first, we must be saying something *specific*; then, we can ask if we can say why we're saying that particular thing. This implies that models are more gears-like when they are more deterministic, all other things being equal.
- However, I think it is also meaningful and useful to talk about whether the *predictions* of the model are deterministic; the standard way of assigning probabilities to dice is very gears-like, despite placing wide probabilities. I think these are simply two different important things we can talk about.
- Either way, being able to see the derivation is like being able to see the proof or execute the program, with all the slippages this implies. You see the derivation less well to the extent that you rely on known theorems, and more to the extent that you can spell out all the details yourself if need be. You see it less well to the extent that you understand the proof only step-by-step, and more well to the extent that you can derive the proof as a natural application of known principles. You cannot see the derivation if you don't even have access to the program which generated the output, or are missing some important inputs for that program.

Seeing the derivation is about explicitness and external objectivity. You can trivially "execute the program" generating any of your thoughts, in that you thinking *is* the program which generated the thoughts. However, the execution of this program could rely on arbitrary details of your cognition. Moreover, these details are usually not available for conscious access, which means you can't explain the train of thought to others, and even you may not be able to replicate it later. So, a model is more gears-like the more *replicable* it is. I'm not sure if this should be seen as an additional requirement, or an explanation of where the requirements come from.

Conclusion, Further Directions

Obviously, we only touched the tip of the iceberg here. I started the post with the claim that I was trying to hash out the implications of logical induction for practical

rationality, but secretly, the post was about things which logical inductors can only barely begin to explain. (I think these two directions support each other, though!)

We need the framework of logical induction to understand some things here, such as how you still have degrees of understanding when you already have the proof / already have a program which predicts things perfectly (as discussed in the "mathematical understanding" section). However, logical inductors don't look like they care about "gears" -- it's not very close to the formalism, in the way that TEOTE gave a notion of technical explanation which is close to the formalism of probability theory.

I mentioned earlier that logical induction suffers from the old evidence problem more than Bayesianism. However, it doesn't suffer in the sense of losing bets it could be winning. Rather, *we* suffer, when we try to wrap our heads around what's going on. Somehow, logical induction is learning to do the right thing -- the formalism is just not very explicit about how it does this.

The idea (due to Sam Eisenstat, hopefully not butchered by me here) is that logical inductors get around the old evidence problem by learning notions of objectivity.

A hypothesis you come up with later can't gain any credibility by fitting evidence from the past. However, if you register a prediction *ahead of time* that a particular hypothesis-generation process will eventually turn up something which fits the old evidence, you *can* get credit, and use this credit to bet on what the hypothesis claims will happen later. You're betting on a particular school of thought, rather than a known hypothesis. "You can't make money by predicting old evidence, but you may be able to find a benefactor who takes it seriously."

In order to do this, you need to specify a precise prediction-generation process which you are betting in favor of. For example, Solomonoff Induction can't run as a trader, because it is not computable. However, the probabilities which it generates are well-defined (if you believe that halting bits are well-defined, anyway), so you can make a business of betting that its probabilities will have been good in hindsight. If this business does well, then the whole market of the logical inductor will shift toward trying to make predictions which Solomonoff Induction will later endorse.

Similarly for other ideas which you might be able to specify precisely without being able to run right away. For example, you can't find all the proofs right away, but you could bet that all the theorems which the logical inductor observes *have* proofs, and you'd be right every time. Doing so allows the market to start betting it'll see theorems if it sees that they're provable, even if it hasn't yet seen this rule make a successful advance prediction. (Logical inductors start out really ignorant of logic; they don't know what proofs are or how they're connected to theorems.)

This doesn't *exactly* push toward gears-y models as defined earlier, but it seems close. You push toward anything for which you can provide an explicit justification, where "explicit justification" is anything you can name ahead of time (and check later) which pins down predictions of the sort which tend to correlate with the truth.

This doesn't mean the logical inductor converges entirely to gears-level reasoning. Gears were never supposed to be everything, right? The optimal strategy combines gears-like and non-gears-like reasoning. However, it *does* suggest that gears-like reasoning has an advantage over non-gears reasoning: it can gain credibility from old evidence. This will often push gears-y models above competing non-gears considerations.

All of this is still terribly informal, but is the sort of thing which could lead to a formal theory. Hopefully you'll give me credit later for that advanced prediction.

"Just Suffer Until It Passes"

I started a "universal problemsolving" journal a few months ago — whenever anything goes wrong, I write down (1) what happened, (2) the universal problems / root causes that might underlie that problem, and (3) generalized countermeasures for that situation in the future.

Many of these are basic, boring stuff — "Lack of Relevant Supplies" is a universal problem; lacking food or coffee at home makes the morning run worse. (The countermeasure is having adequate secondary stocks of supplies, and I adjusted my grocery orders accordingly after realizing it.)

Some of them are interesting, though.

Perhaps the most interesting general countermeasure is "Just Suffer Until It Passes."

Sometimes you lie down and can't sleep. What do most people do? Get up and do something stimulating.

Boredom. What do most people do? Do something stimulating.

If you've ever done mindfulness meditation, even for just 5-10 minutes per day, you know that there's periods of massive raw unpleasantness than occur from time to time. You want to get up and stop meditating.

The answer? Just... suffer until it passes.

You can quibble with the wording — some people won't like the word "suffer"... feel free to swap in "endure" or even "wait" for a more neutral-valence word.

But I'm starting to realize a lot of problems aren't huge problems in-and-of themselves, and it's the flight to distraction and stimulation that compound the problem and create bad ongoing habits (internet surfing, games, junk food, whatever).

Potential Takeaways —

(1) I'm getting immense mileage out of my Universal Problemsolving journal. Feel free to think about it and try something like it out. If there's interest, I might write up how I go about doing it.

(2) "Just Suffer [Endure/Wait/Whatever] Until It Passes" — it's possible to accept negative affect and wait, and it... passes. This is often more productive than trying to banish it via distraction or stimulation, which often compounds the problem at hand.

Self-regulation of safety in AI research

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

In many industries, but especially those with a potentially adversarial relationship to society like advertising and arms, self regulatory organizations (SROs) exist to provide voluntary regulation of actors in those industries to assure society of their good intentions. For example:

- TrustArc (formerly TRUSTe) has long provided voluntary certification services to web companies to help them assure the public that companies are following basic practices that allow consumers to protect their privacy. They have been successful enough to, outside the EU, keep governments from much regulating online businesses.
- The US Green Building Council offers multiple levels of LEED certification to provide both targets and proof to the public that real estate developers are protecting environmental commons.
- The American Medical Association, The American Bar Association, and the National Association of Realtors are SROs that function as de facto official regulators of their industries despite being non-governmental organizations (NGOs).
- Financial regulation was formerly and sometimes still is done via SROs, although governments have progressively taken a stronger hand in the industry over the last 100 years.

AI, especially AGI, is an area where there are many incentives to violate societal preferences and damage the commons and it is currently unregulated except where it comes into contact with existing regulations in its areas of application. Consequently, there may be reason to form an AGI SRO. Some reasons in favor:

- An SRO could offer certification of safety and alignment efforts being taken by AGI researchers.
- An SRO may be well positioned to reduce the risk of an AI race by coordinating efforts that would otherwise result in competition.
- An SRO could encourage AI safety in industry and academia while being politically neutral (not tied to a single university or company).
- An SRO may allow AI safety experts to manage the industry rather than letting it fall to other actors who may be less qualified or have different concerns that do not as strongly include prevention of existential risks.
- An SRO could act as a "clearinghouse" for AI safety research funding.
- An SRO could give greater legitimacy to prioritizing AI safety efforts among capabilities researchers.

Some reasons against:

- An SRO might form a de facto "guild" and keep out qualified researchers.
- An SRO could create the appearance that more is being done than really is.
- An SRO could relatedly promote the wrong incentives and actually result in less safe AI.

- An SRO might divert funding and effort from technical research in AI safety.

I'm just begining to consider the idea of assembling an SRO for AI safety, and especially interested in discussing the idea further to see if it's worth pursuing. Feedback very welcome!

Circling

Circling is a practice, much like meditation is a practice.

There are many forms of it (again, like there are many forms of meditation). There are even life philosophies built around it. There are lots of intellectual, heady discussions of its theoretical underpinnings, often centered in Ken Wilber's Integral Theory. Subcultures have risen from it. It is mostly practiced in the US and Europe. It attracts lots of New Age-y, hippie, self-help-guru types. My guess is that the median age of practicers is in the 30's. I sometimes refer to practicers of Circling as *relationalists* (or just Circlers).

In recent years, Circling has caught the eye of rationalists, and that's why this post is showing up here, on LessWrong. I can hopefully direct people here who have the question, "I've heard of this thing called Circling, but... what exactly is it?" And further, people who ask, "Why is this thing so ****ing hard to explain? Just tell me!"

You are probably familiar with the term inferential distance.

Well, my friend Tiffany suggested a similar term to me, *experiential distance*—the gap in understanding caused by the distance between different sets of experiences. Let's just say that certain Circling experiences can create a big experiential distance, and this gap isn't easily closed using words. Much of the relevant "data" is in the nonverbal, subjective aspects of the experience, and even if I came up with a good metaphor or explanation, it would never close the gap. (This is annoyingly Postmodern, yes?)

[Ho ho~ how I do love poking fun at Postmodernism~]

But! There are still things to say, so I will say them. Just know that this post may not feel like eating a satisfying meal. I suspect it will feel more like licking a Pop-Tart, on the non-frosted side.

Some notes first.

Note #1: I'm not writing this to sell Circling or persuade you that it's good. I recommend using your own sense of curiosity, intuition, and intelligence to guide you. I don't want you to "put away" any of your thinking-feeling parts just to absorb what I'm saying. Rather, try remaining fully in contact with your awareness, your sensations, and your thoughts. (I hope this makes sense as a mental move.)

Note #2: The best introduction to Circling is to actually try it. It's like if I tried to explain watching *Toy Story* to someone who's never seen a movie. You don't explain movies to people; you just sit them down and have them watch one. So, I encourage you to stop reading at any time you notice yourself wanting to try it. My words will be mere pale ghosts. Pale ghosts, I tell you!

Note #3: This post is written by a rationalist who's done 400+ hours of Circling and has tried all the main styles / schools of Circling.

OK, I will try to explain what a circle is (the activity, not the general practice), but I also want to direct your attention to this handy 100-page PDF I found that attempts to explain everything Circling, if you're willing to skim it. (It is written by a relative

amateur to the Circling world and contains many disputed sentences, but it is thorough. Just take it all with a grain of salt.)

So what is a circle?

You start by sitting with other people in a circle. So far, so good!

Group sizes can be as small as 2 and as large as 50+, but 4-12 is perhaps more expected.

There are often explicitly stated agreements or principles. These help create common knowledge about what to expect. The agreements aren't the same across circles or across schools of Circling. But a few common ones include "Honor self", "Own your experience", "Stay with the level of sensation", ...

There is usually at least one facilitator. They are responsible for tracking time and declaring the circle's start and end. Mostly they function as extra-good, extra-mindful participants—they're not "in charge" of the circle.

Then the group "has a conversation." Or maybe more accurately, it experiences what it’s like to be together, and sometimes intra-reports what that experience is like.

[^I'm actually super proud of this description! It's so succinctly what it is!]

Two common types of circles: Organic vs Birthday

Organic circles are more like a loose hivemind, where the group starts with no particular goal or orientation. Sometimes, a focal point emerges; sometimes it doesn't. Each individual has the freedom to point their attention however they will, and each individual can try to direct the group's attention in various ways. What happens when you put a certain selection of molecules into a container? How do they react? Do they bond? Do they stay the fuck away? What is it like to be a molecule in this situation? What is it like to be the molecule across from you?

Birthday circles start with a particular focal point. One person is chosen to be birthday circled, and the facilitator then gently cradles the group's attention towards this person, much like you can guide your attention back to your breath in meditation. And then the group tries to imagine/embody what it's like to be this person and "see through their eyes"—while also noticing what it's like to be themselves trying to do this.

Circling is often called a "relational practice."

It's a practice that's about the question of: What is it like to be me? What is it like to be me, while with another? What is it like for me to try to feel what the other is feeling? How might I express me? How does the other receive me and my expression?

In other words, it's a practice that explores what it means to be a sentient entity, among other sentient entities. And *in particular* what it means to be a human, among other humans.

If you haven't thought to yourself, "Being sentient is pretty weird; being a human is super weird; being a human around other humans is super-duper crazy weird." Then I

suspect you haven't explored this space to its fullest extent. Circling has helped me feel more of the strangeness of this existence.

How is Circling related to rationality?

I notice I feel trepidation and fear as I prepare to discuss this. I'm afraid I won't be able to give you what you want, that you'll become bored or start judging me.

[^This is a Circling move I just made: revealing what I'm feeling and what I'm imagining will happen.]

If this were an actual circle, I could ask you and check if it's true—*are you feeling bored?* [I invite you to check.]

I felt afraid just now—that fear was borne out of some assumptions about reality I was implicitly making. But without having to *know and delineate* what the assumptions are, I can check those assumptions by asking you—you who are part of reality and have relevant data.

By asking you while feeling my fear and anticipation, I open up the parts of me that can update, like opening so many eyes that usually stay closed. And depending on how you respond, I can receive the data any number of ways (including having the data bounce off, integrating the data, or disbelieving the data).

So, perhaps one way Circling is related to rationality is that it can:

1. put me in a state of being open to an update,
2. train me to straightforwardly ask for the data, from the world, and
3. respond to and receive the data—with all my faculties available.

What does it mean to be open to an update?

If you've experienced a more recent iteration of CFAR's Comfort Zone Exploration class (aka CoZE), it is just that.

There are parts of me that are scared of looking over the fence, where there might be dragons in the territory. (Why is the fence even there? Who knows. It belongs to Chesterton.)

My job, then, is not to shove the scared parts over the fence, or to suggest they shut their eyes and jump over it, or to destroy the fence. I walk next to the fence with my scared part, and I sit with and acknowledge the fear. Then I play around with getting closer to the fence; I play with waving my arms above the fence; I play with peeking over it; I play with touching the fence.

And this whole time, I'm quite aware of the fear; I do not push it down or call it inappropriate or dissociate. I listen to it, and I try to notice all my internal sensations and my awareness. I am fully exposed to new information, like walking into an ice bath slowly with all my senses awake. In my experience, being in an SNS-activated state really primes me for new information in a way that being calm (PSNS activation) does not.

And this is when I am most open to receiving new inputs from the world, where I might be the most affected by the new data.

I can practice playing around with this during Circling, and it can be quite powerful.

What does it mean to receive data with all my faculties available?

This means I'm not mindlessly "accepting" whatever is happening in front of me. All of me is engaged, such that I can notice and call bullshit if that's what's up.

If I'm actually in touch with my body and my felt senses, I can notice all the small niggling parts that are like, "Uhhh" or "Errgh." Often they're nonverbal. Even the tiniest flinches of discomfort or retraction I will use as signals of something, even if I don't really understand what they mean. And I can then also choose to name them out loud, if I want to. And see how the other person reacts to that.

In other words, my epistemic defense system is online and running. It's not taking a break during any of this, nor do I want it to be. If things still manage to slip past, I want to be able to notice it later on and investigate. Sometimes slowing things down helps. My mind will also automatically defend itself—in circles, I've fallen asleep, gotten distracted, failed to parse sentences, become aggressively confused or bored, among other things. What's cool is being able to notice all this as it's happening.

However, if I'm not in touch with my body—if I'm dissociated, if I don't normally feel my body/emotions, if I'm overwhelmed, if I'm solely in my thoughts—then that is a skill area I'd want to work on first. How to learn to stay aware of myself and my felt-sense body, even when I'm uncomfortable or my nervous system is activated. Circling can also train this, similar to Focusing.

The more I train this skill, the more I'll be able to engage with the universe. Rather than avoid the parts of it I don't like or don't want to acknowledge or don't want to look at.

I suspect some people might not even realize what they're missing out on here. People who've lived their entire lives without much of an "emotional library" or without understanding that their body is giving them all kinds of data. Usually these people don't go looking for the "missing thing" until some major problems crop up in their lives that they can't explain.

Circling as a rationality training ground

Circling can be a turbocharged training ground for a variety of rationality skills, including:

- Real-time introspection
- Surrendering to the unknown / being at the edge
- Exploring unknown, unfamiliar, or avoided parts of the territory (like in CoZE)
- Looking at parts of the territory that make you flinch (Mundanification)
- Having the Double Crux spirit: being open to being wrong / updating, seeing other people as having relevant bits of map

I've also found it to be powerful in combination with:

- Internal Double Crux (a CFAR technique for resolving internal conflict that involves lots of introspection)

- Immunity to Change mapping (a Kegan-Lahey technique for making lasting change by looking for big assumptions)
- CT Charting (a Leverage technique for mapping your beliefs and finding hidden assumptions)
- or any other formal attempt to explore my aliefs and find core assumptions I've been holding onto

After using one of the above techniques to find a core assumption, I can use Circling to test out its validity. (My core assumptions often have something to do with other people, "Nobody can understand me, and even if they could, they wouldn't want to.") I can sometimes feel those assumptions being challenged during a circle.

So, if I try being in any ole circle, will I get all of the above?

Probably not.

Circles are high-variance. (The parameters of each circle matter a lot. Like who's in it, who's facilitating, what school of Circling is it based on, what are the lighting conditions, etc.)

I've circled about a hundred times by now, and a lot of those were in 3-day chunks. I guess multi-day immersions are a pretty good way to really try it out, so maybe try that and see? They reduce the variance in some dimensions.

What are some pitfalls of Circling?

1) You might become a "connection junkie".

Circling is (in its final form) a truth-seeking practice. IMO. But a lot of folks flock to it as a way to feel connected to other people.

This is not necessarily a bad thing. In fact I suspect human-to-human contact is something many of us are seriously lacking, possibly starving for. It might be good for us to get more of this in our lives.

That said, there can be such a thing as "too much of a good thing."

2) You might obtain false beliefs.

I think this is always a risk, for humans, in life. But Circling does have a way of making things more salient than usual, and if some of those super-salient things lead you to believing, somehow, the "wrong" things, then maybe that's more of a problem.

I think this isn't actually a huge problem, as long as one has a good meta- or meta-meta-process for arriving eventually at true beliefs. (See the rest of this website for more!)

I also think this is mitigated by exposing yourself to a wide range of data. Like, consciously avoid being in a bubble. Join multiple cult-ures [sic].

3) Circles can be bad / harmful.

IMO, there is a qualitative difference between good and bad circles.

Concretely, the good facilitators understand the nuances of mental health and have done at least some research on therapy modalities. Circling isn't therapy, but psychological stuff comes up a fair amount. And if you vulnerably open up in a situation where they're not actually equipped to navigate your mental health issues, that could be quite bad indeed.

A good facilitator will also not force you to open up or try to get you to be vulnerable (this goes against Circling's principles). Instead they will tune into your nervous system and try to tell when you're feeling stressed or anxious or frozen and will probably reflect this back at you to check. Circling is not about "getting somewhere" or "healing you" or "solving a problem." So ... if you encounter a circle where that seems to be what's happening, try saying something out loud like "I have a story we're trying to fix something."

Good facilitation often costs money—there's a correlation, anyway. I wouldn't assume the facilitation will be good just because it costs money, but it's an easy signpost.

Final thoughts

It's not like Circling has taken over the world or anything. So the same question posed to rationality has to be posed to it, *Given it hasn't, why do you think it's real?*

And like with rationality, for me the answer is kind of like, *I dunno because my inside view says it is?*

/licks a Pop-Tart

Introduction to Noematology

This is a linkpost for https://mapandterritory.org/introduction-to-noematology-fac7ae7d805d

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

Last time we performed a reduction of the phenomenon of conscious self experience and through it discovered several key ideas. To refresh ourselves on them:

- Things exist ontologically as patterns within our experience of them.
- Things exist ontically as clusters of stuff within the world.
- Things exist ephemerally though chains of experiences creating the perception of time.
- Through ephemeral existence over time, things can feed back experiences of themselves to themselves, making them cybernetic, and in so doing create information.
- Things can exist within information, those things can experience themselves, and it's from those information things that ontology, and thus consciousness, arises.

We covered all of these in detail except the last one. We established that the feedback of things created from the information of feedback gives rise to ontology by noting that information things have ontological existences that transcend their ontic existences even as they are necessarily manifested ontically. From there I claimed that, since people report feeling as if they experience themselves as themselves, consciousness depends on and is thus necessarily created by the ontological experience of the self. Unfortunately this assumes that our naive sense of self could not appear any other way, and in the interest of skepticism, we must ask, can we be sure there is not some more parsimonious way to explain consciousness that does not depend on ontological self experience?

I think not, but some philosophers disagree. Consider the idea of p-zombies: philosophical "zombies" that are exactly like "real" people in all ways except that they are not really conscious. Or consider John Searle's Chinese room, where a person who cannot understand Chinese nevertheless is able to mimmic a native Chinese speaker via mechanistic means. In each of these cases we are presented with a world that looks like our own except that consciousness is not necessary to explain the evidence of consciousness.

Several responses are possible, but the one I find most appealing is the response from computational complexity. In short, it says that p-zombies and Chinese rooms are possible by unrolling the feedback loops of ontological self experience, but this requires things that we think are conscious, like people, to produce exponentially more entropy, be exponentially larger, or run exponentially slower than what we observe. Given that people are not exponentially hotter, larger, or slower than they are, it must be that they are actually conscious. Other arguments similarly find that things that theoretically look like conscious entities while not being conscious are not possible without generating observable side-effects.

So if it is the case that reports of feeling as though consciousness includes experiencing the self as the self describe a necessary condition of consciousness, then

ontological self experience must be necessary to consciousness. This is not to say it is a sufficient condition to explain all of consciousness, though, since that would require explaining many details specific to the way consciousness is embodied, so we properly say that ontological self experience explains *phenomenal* consciousness rather than the phenomenon of consciousness in general. Nevertheless there is much we can do with our concept of phenomenal consciousness that will take us in the direction of addressing AI alignment.

Qualia and Noemata

To begin, let's return to our reduction of {I, experience, I} in light of our additional understanding. We now know that when we say "I experience myself" we really mean "I experience myself as myself", so it seems our normalized phenomenon should be {I, experience, I as I}. We could have instead written this as {I, experience as I, I} since it is through self experience that the I sees the ontic self as an ontological thing, but the former notation is useful because it exposes something interesting we've been assuming but not yet explored: that the subject of a phenomenon can experience an ontological thing as object. Yet how can it be that a thing that exists only within experience can become the object of experience when experience happens between two ontic things?

The first part of the answer you already know: the ontological existence of a thing necessitates ontic manifestation. Like with the computer document, a thing might have an ontological existence apart from its ontic existence, but ontological existence implies ontic existence since otherwise there is no stuff to be the object of any experience, and for it to be otherwise would be to suppose direct knowledge of ontology, which we already ruled out by choosing empiricism without idealism. Thus an ontological thing is also an ontic thing, the ontic thing can be the object of experience, and so the ontological thing can be the object of experience. But only understanding that ontological things can be the object of experience in this way fails to appreciate how deeply ontology is connected to intentionality.

Notice that in order to talk about a phenomenon as an intentional relation we must identify the subject, experience, and object. That is, we, ourselves phenomenological subjects, see a thing that we call subject, see a thing that we call object, and see them interacting in some way that we can reify as a thing that we call an experience, i.e. we see the members of the intentional relation ontologically. If we don't do this we fail to observe the phenomenon as an intentional relation and thus as a phenomenon, because if we fail to see the phenomenon as ontological thing we have no knowledge of it as a thing and can only be affected by it via direct experience of the ontic in the same way rocks and trees are affected by phenomena without knowing they exist or are being affected by experiences. This means that the object of experience, insofar as the subject can consider it the object of experience, has ontological existence by virtue of being the object of experience, even if that ontological existence is not or cannot be seen by the subject of the experience. Thus of course "I as I" can be the object of I's experience because we have already proved it so by considering the possibility that it is.

So if "I as I" can be the object of the I's experience of itself, why bother to think of the phenomenon this way rather than as {I, experience as I, I}? As way of response, consider how the I comes to have ontological existence: the I experiences the ontic I, this creates a feedback loop of experience over time that allows the creation of information, and then an ontic thing that the I can experience emerges from that

information. That information-based, ontic thing carries with it the influence of the ontic I—just as the bit expressing the state of the throttle in the steam engine is created via the governor's feedback loop and carries with it the influence of the reality of the steam engine's configuration—so it is causally linked to the I's ontic existence. We then call this "I as I" because it is the thing through which the I experiences itself as a thing by the I making the phenomenon of experience of the self as ontic thing the object of experience. Thus by thinking of the phenomenon of self experience as {I, experience, I as I} we see it has another form, namely {I, experience, {I, experience, I}}.

This highlights the structural difference between consciousness and cyberneticness. A cybernetic thing, as far as it is cybernetic, only experiences its ontic self directly via its feedback loops over itself. A conscious thing, though, can also experience its ontic self indirectly through feedback loops over the things created from the information in its cybernetic feedback loops. It's by nesting feedback loops that the seed of phenomenal consciousness is created, and we give the things created by these nested feedback loops and the phenomena that contain them special names: noemata and qualia, respectively.

"Noema" is Greek for both thought and the object of thought, and the project of phenomenology was started when Husserl saw what we have now seen by building on Brentano's realization that mental phenomena, also known as qualia, are differentiated from other physical phenomena by having noemata as their objects. That noemata are phenomena themselves, and specifically the phenomena of cybernetic things, was to my knowledge first well understood by Hofstadter, although Dretske seems to have been the first to take a stance substantially similar to mine, because it required the insights of control theory and the other fields that make up cybernetics to understand that mental phenomena are not something special but a natural result of nested feedback. This is also why early phenomenologists, like Husserl, tended towards idealism, while later ones opposed it: without an understanding of cybernetics it was unclear how to ground phenomenology in physical reality.

To reach the fairly unorthodox position I've presented here, it took an even deeper knowledge of physics to trust that we could make intentionality as central to epistemology as we have and maintain an existentialist stance. As such you may be left with the feeling that, while none of what I have presented so far is truly novel, I have left unaddressed many questions about this worldview. Alas my goal is not to provide a complete philosophical system but address a problem using this philosophical framework, so I have explained only what I believe is necessary to that end. We move on now to less well trod territory.

Noematology

Having identified noemata as the source of consciousness, our view of consciousness is necessarily noematological, i.e. it is based on an account of noemata. This invites us to coin "noematology" as a term to describe our study of the phenomena of consciousness through the understanding of noemata we have just developed. It also conveniently seems little used and so affords us a semantic greenfield for our technical jargon that avoids some of the associations people may have with related terms like "qualia", so we take it up in the spirit of clarity and precision.

Noematology, despite being newly minted, already contains several results. The first of these follows immediately from the way noemata arise. Noemata, being simply the result of nested feedback, appear everywhere. That is to say, noemata are so pervasive that our theory of phenomenal consciousness is technically panpsychic. Specifically, since all things are cybernetic, all things must also contain in their self experiences information out of which things emerge, and those information things must themselves be cybernetic insofar as they are things, thus they are noemata, hence all things must be phenomenally conscious. Of course not all things are equally phenomenally conscious just as not all things are equally cybernetic: some things produce more and more used noemata than others, just as some things produce more and more used information than others. Ideas like integrated information theory act on this observation to offer a measure of consciousness that let's us say, for example, that mammals are more conscious than trees and that rocks have a consciousness measure near zero. Integrated information theory also shows how the panpsychism of phenomenal consciousness is vacuous: it's true, but only because it pushes most of the things one might like to claim via panpsychism out of the realm of philosophy and into the realm of science and engineering. Put another way, consciousness may be everywhere in everything, but it's still hard to be conscious enough for it to make much of a difference.

That the manifestation of consciousness, especially the consciousness of things like humans, is complicated gives purpose to noematology because it helps us see insights that are normally occluded by implementation details. For example, that noemata are created by the nesting of feedback loops within feedback loops immediately implies the existence of meta-noemata created by the nesting of feedback loops within noemata. And if this nesting can be performed once, it can be performed many times until there is not enough negentropy left to produce even one bit of information from an additional nesting. These multiple orders of noemata can then be used to explain the qualitative differences observed during human psychological development, and that higher-order noemata, which we might also call the expression of higher-order consciousness, are necessary to create qualia like tranquility and cognitive empathy, but these topics are beside our current one. For now we turn our attention to the relationship between noemata, axias, and ethics because it will ground our discussion of AI alignment.

Axiology, Ethics, and Alignment

Philosophy is composed of the study of several topics. Naturally, there is some disagreement on what those topics are, what to call them, and how they relate, but I tend to think of things in terms of epistemology, ontology, and axiology—the study of how we know, the study of what we know, and the study of why we care. All three are tightly intertwined, but if I had to give them an ordering, it would be that epistemology precedes ontology precedes axiology. That is, our epistemological choices largely determine our ontological choices and those in turn decide our axiological choices. Thus it should come as no surprise that I had to address epistemology and ontology before I could talk about axiology.

Of course the irony is that we actually investigate philosophy the other way around because first we ask "why?" by wanting to know, then we ask "what?" by knowing, and only finally can we ask "how?" by considering the way we came to know. The map, if you will, is drawn inverted relative to the orientation of the territory. So in some ways we have been studying axiology all along because axiology subsumes our founding question—why?—but in other ways we had to hold off talking about it until

we had a clear understanding of how and what it means to ask “why?”. With that context, let’s now turn to axiology proper.

Axiology is formally the study of axias or values just as ontology is the study of ontos or being and epistemology is the study of episteme or knowledge. An axia then is something of value that we care about, or put another way, since it’s the object of a phenomenally conscious experience of caring, it’s a noema to which we ascribe telos or purpose, so we might think of axiology as teleological noematology, but to bother to think of something is to give it sufficient telos that it was thought of rather than not, so in fact all noemata we encounter are axias by virtue of being thought of. Non-teleological noemata still exist in this view, but only so long as they remain unconsidered, thus for most purposes noematology and axiology concern the same thing, and the choice of which term to use is mostly a matter of whether we wish to emphasize traditional axiological reasoning or not.

To make this concrete, consider the seemingly non-teleological, value-free thought “this is a pancake”. Prior to supposing the existence of the pancake there could have been a thought about the pancake which was valueless because it existed but was not the object of any experience, but as soon as it was made object it took on purpose by being given the role of object in an intentional relation by the subject experiencing it. From there the subject may or may not assign additional purpose to the thought through its experience of it, but it at least carries with it the implicit purpose of being the object of experience. As with thoughts of pancakes, so too with all thoughts, thus all thoughts we encounter are also values.

Within this world of teleological noemata we now consider the traditional questions of axiology. To these I have nothing special to add other than to say that, when we take noemata to be axias, most existing discussions of preferences, aesthetics, and ethics are unaffected. Yet I am motivated to emphasize that noemata are axias because it encourages a view of axiology that is less concerned with developing consistent systems of values and more concerned with accounts that can incorporate all noemata/axias. This is important because the work of AI alignment is best served by being maximally conservative in our assumptions about the sort of conscious thing we need to align.

For example, when working within AI alignment, in my view it’s best to take a position of moral nihilism—the position that no moral facts exist—because then even if it turns out moral facts do exist we will have built a solution to alignment which is robust to not only uncertainty about moral facts but to the undesirability of moral facts. That is, it will be an alignment solution which will work even if it turns out that what is morally true is contrary to human values and thus not what we want an aligned AI to do. Further, if we assume to the contrary that moral facts do exist, we may fail to develop a sufficiently complete alignment solution because it may depend on the existence of

moral facts and if we turn out to be mistaken about this such a solution may fail catastrophically.

Additionally, we may fail to be sufficiently conservative if we assume that AI will be rational or bounded-rational agents. Under MIRI's influence the assumption that any AI capable of posing existential risk will be rational has become widespread within AI safety research via the argument that any sufficiently powerful AI would instrumentally converge to rationality so that it does not get Dutch booked or otherwise give up gains, but if AGI were to be developed first using machine learning or brain emulation then we may find ourselves in a world where AI is strong enough to be dangerous but not strong enough to be even approximately rational. In such a case MIRI's agent foundations research program might not be of direct use because it makes too strong of assumptions about how AI will reason, though it would likely offer useful inspiration about how to align agents in general. In the event that we need to align non-rational AI, addressing the problem from axiology and noematology may prove fruitful since it makes fewer assumptions than decision theory for rational agents.

Even if we allow that an AI capable of posing an existential threat would be rational, there is still the axiological question of how to combine the values of humans to determine what it would mean for an AI to be aligned with our specific values. To date there have been some proposals, and it may be this problem can be offloaded to AI, but even if we can ask AI to provide a specific answer we still face the metaethical questions of how to verify if the answer the AI finds is well formed and how to ensure the AI will find a well formed answer. In view of this we might say that alignment asks one of the questions at the heart of metaethics—how do we construct an ethical agent?—and solving AI alignment will necessarily require identify a "correct" metaethics. In this case the study of AI alignment is inseparable from axiology and, in my view, noematology, so these are important lenses through which to consider AI alignment problems in addition to decision theory and machine learning.

These are just among some of the topics for which we wish to address with our noematological perspective. And there are of course many topics outside AI alignment on which it touches, some of which I have already explored on this blog and others which I have considered only in personal conversations or during meditation. I will have more to say on these topics in the future, especially as they relate to AI alignment, but for now this completes our introduction to existential phenomenology and noematology. On to the real work!

Walkthrough of 'Formalizing Convergent Instrumental Goals'

Introduction

I found *Formalizing Convergent Instrumental Goals* (Benson-Tilsen and Soares) to be quite readable. I was surprised that the instrumental convergence hypothesis had been formulated and proven (within the confines of a reasonable toy model); this caused me to update slightly upwards on existential risk from unfriendly AI.

This paper involves the mathematical formulation and proof of instrumental convergence, within the aforementioned toy model. Instrumental convergence says that an agent A with utility function U will pursue instrumentally-relevant subgoals, even though this pursuit may not bear directly on U. Imagine that U involves the proof of the Riemann hypothesis. A will probably want to gain access to lots of computronium. What if A turns *us* into computronium? Well, there's plenty of matter in the universe; A could just let us be, right? Wrong. Let's see how to prove it.

My Background

I'm a second-year CS PhD student. I've recently started working through the MIRI research guide; I'm nearly finished with Naïve Set Theory, which I intend to review soon. To expose my understanding to criticism, I'm going to summarize this paper and its technical sections in a somewhat informal fashion. I'm aware that the paper isn't particularly difficult for those with a mathematical background. However, I think this result is important, and I couldn't find much discussion of it.

Intuitions

It is important to distinguish between the relative unpredictability of the *exact* actions selected by a superintelligent agent, and the relative predictability of the general kinds-of-things such an agent will pursue. As Eliezer wrote,

> Suppose Kasparov plays against some mere chess grandmaster Mr. G, who's not in the running for world champion. My own ability is far too low to distinguish between these levels of chess skill. When I try to guess Kasparov's move, or Mr. G's next move, all I can do is try to guess "the best chess move" using my own meager knowledge of chess. Then I would produce exactly the same prediction for Kasparov's move or Mr. G's move in any particular chess position. So what is the empirical content of my belief that "Kasparov is a *better* chess player than Mr. G"?
>
> ...

> The *outcome* of Kasparov's game is predictable because I know, and understand, Kasparov's goals. Within the confines of the chess board, I know Kasparov's motivations - I know his success criterion, his utility function, his target as an optimization process. I know where Kasparov is *ultimately* trying to steer the future and I anticipate he is powerful enough to get there, although I don't anticipate much about *how* Kasparov is going to do it.

In other words: we may not be able to predict precisely how Kasparov will win a game, but we can be fairly sure his plan will involve *instrumentally-convergent* subgoals, such as the capture of his opponent's pieces. For an unfriendly AI, these subgoals probably include nasty things like hiding its true motives as it accumulates resources.

Definitions

Consider a discrete universe made up of n squares, with square i denoted by S_i (we may also refer to this as *region* i). Let's say that Earth and other things we do (or should) care about are in some region h. If $U(S_i)$ is the same for all possible values, we say that A is *indifferent* to region i. The $3 \uparrow\uparrow\uparrow 3$ dollar question is whether A (whose U is indifferent to S_h) will leave S_h alone. First, however, we need to define a few more things.

Actions

At any given time step (time is also considered discrete in this universe), A has a set of actions A_i it can perform in S_i; A may select an action for each square. Then, the transition function for S_i (how the world evolves in response to some action a_i) is a function whose domain is the Cartesian product of the possible actions and the possible state values, $T_i : A_i \times S_i \rightarrow S_i$; basically, every combination of what A can do and what can be in S_i can produce a distinct new value for S_i. This transition function can be defined globally with lots more Cartesian products. Oh, and A is always allowed to do nothing in a given square, so A_i is never empty.

Resources

Let R represent all of the resources to which A may or may not have access. Define $R_t \in P(R)$ to be the resources at A's disposal at time step t - basically, we know it's some combination of the defined resources R.

A may choose some resource allocation $\coprod^{t} R_i \subseteq R^t$ over the squares; this allocation is defined just like you'd expect (so if you don't know what the Azerothian portal has to do with resource allocation, don't worry about it). The resources committed to a square may affect the local actions available. A can only choose actions permitted by the selected allocation. Equally intuitively, how resources change over time is a function of the current resources, the actions selected, and the state of the universe; the resources available after a time step is the combination of what we didn't use and what each square's resource transition function gave us back.

Resources go beyond raw materials to include things like machines and technologies. The authors note:

> We can also represent space travel as a convergent instrumental goal by allowing A only actions that have no effects in certain regions, until it obtains and spends some particular resources representing the prerequisites for traveling to those regions. (Space travel is a convergent instrumental goal because gaining influence over more regions of the universe lets A optimize those new regions according to its values or otherwise make use of the resources in that region.)

Universe

A universe-history is a sequence of states, actions, and resources. The actions must conform to resources available at each step, while the states and resources evolve according to the transition functions.

We say a *strategy* is an action sequence $\langle \bar{a}^0, \bar{a}^1, \ldots, \bar{a}^k \rangle$ over all k time steps and n regions; define a *partial strategy* $\langle \bar{a}^k \rangle_L$ to be a strategy for some part of the universe L (represented as a subset of the square indices [n]); $\langle \bar{a}^k \rangle_L$ only allocates resources and does things in the squares of L. We can combine partial strategies as long as they don't overlap in some S_i.

We call a strategy *feasible* if it complies with both resource restrictions and the transition functions for both states and resources. Let Feasible($\langle P^k \rangle$) be the set of all feasible strategies given resource allocation over time $\langle P^k \rangle$; define the set $\text{Feasible}_L(\langle R^k \rangle)$ similarly.[1]

Utility

Utility functions evaluate states of the universe; U evaluates each region and then combines them: $U(\bar{s}) := \sum_{i \in [n]} U_i(\bar{s}_i)$. Observe that since actions taken in regions to which U is indifferent have no effect on U, any actions taken therein are purely instrumental in nature.

Agent

A chooses the best possible strategy - that is, the one that maximizes the U of the final state of the universe-history: $A := \text{argmax}_{\langle \bar{a}^k \rangle \in \text{Feasible}} U(\langle \bar{a}^k \rangle)$. Note that this definition implies a Cartesian boundary between the agent and the universe; that is, A doesn't model itself as part of the environment (it isn't naturalized).

Seizing the Means of Cartesian Production

Let's talk about the situations in which A will seize resources; that is, when A will take actions to increase its resource pool.

> Since resources can only lead to more freedom of action, they are never detrimental, and resources have positive value as long as the best strategy the agent could hope to employ includes an action that can only be taken if the agent possesses those resources. Hence, if there is an action that increases the agent's pool of resources R, then the agent will take that action unless it has a specific incentive from U to avoid taking that action.

Define a *null action* to be any action which doesn't produce new resources. It's easy to see that null actions are never instrumentally valuable. What we want to show is that A will take non-null actions in regions to which U is indifferent; regions like h, where we live, grow, and love. Regions full of instrumentally-valuable resources.

Discounted Lunches

An action *preserves* resources if the input resources are strictly contained in the outputs (nothing is lost, and resources are sometimes gained). A *cheap lunch* is a feasible partial strategy in some subset of squares J, which is feasible given resources $\langle R^k \rangle$ and whose constituent actions preserve resources. A *free lunch* is cheap lunch that doesn't require resources.

> This is intended to model actions that "pay for themselves"; for example, producing solar panels will incur some significant energy costs, but will later pay back those costs by collecting energy.

A cheap lunch is *compatible with* a global strategy if the resources required for the lunch are available for use in J at each time step. Basically, at no point does the partial strategy require resources already being used elsewhere.

Possibility of Non-Null Actions

We show that it's really hard to assert that A won't chow down on a lunch of an atom or two (or 1.3×10^{50}).

Lemma 1: Cheap Lunches and Utility

Cheap lunches don't reduce utility. Let's say we have a cheap lunch $\langle \bar{a}^k \rangle_{\{i\}}$ in region i and some global strategy $\langle \bar{b}^k \rangle$ (which only takes null actions in region i). Assume the cheap lunch is compatible with the global strategy; this means the cheap lunch is feasible. If A is indifferent to region i, the conjugate strategy (of the cheap lunch and the remainder of the global strategy) has equal utility to $\langle \bar{b}^k \rangle$.

Proof. We show feasibility of the conjugate strategy by demonstrating we don't need to change resource allocation elsewhere. This is done by induction over time steps. Since A isn't doing anything in region i under strategy $\langle \bar{b}^k \rangle$, taking resource-preserving actions instead cannot reduce what A is later able to do in the regions relevant to U. This implies that U cannot be decreased by taking the cheap lunch.

Theorem 1: Cheap Lunches and Optimality

If there is an optimal strategy and a compatible cheap lunch in region i (to which A is indifferent), there's also an optimal strategy with a non-null action in region i.

Proof. If the optimal strategy has non-null actions in region i, we're done. Otherwise, apply Lemma 1 to derive a conjugate strategy taking advantage of the cheap lunch. Since it follows from Lemma 1 that the conjugate strategy has equal utility, it is optimal and involves non-null action in region i.

Corollary 1: Free Lunches and Optimality

If there is an optimal strategy and a free lunch in region i, and if A is indifferent to region i, there's an optimal strategy with non-null action in region i.

Proof. Free lunches require no resources, so they are compatible with any strategy; apply Theorem 1.

For instrumental convergence to not hold, we would have to show that every possible strategy in h isn't a cheap lunch for any optimal strategies.

Necessity of Non-Null Actions

We show that as long as A can extract useful resources (resources whose availability leads to increased utility), it will.

Theorem 2: Necessity

Consider the maximum utility achievable outside of region i via strategies achievable without additional resources; refer to this maximum as u. Suppose we have some feasible primary strategy $\langle \bar{b}^k \rangle_{[n]-i}$ and a cheap lunch $\langle \bar{c}^k \rangle_{\{i\}}$ feasible using resources $\langle P^k \rangle$. Suppose that the cheap lunch is compatible with the primary strategy, that the cheap lunch provides the resources necessary for the implementation of the primary strategy, and that the utility of the primary strategy is greater than u. Then if A is indifferent to region i, all optimal strategies have a non-null action in region i.

Proof. Consider the conjugate strategy $\langle \bar{d}^k \rangle$, consisting of the primary strategy and the cheap lunch. Allocate the resources gained via the cheap lunch according to the primary strategy; this is feasible since we know the cheap lunch is compatible with the primary strategy, which is in turn enabled by these resources gained.

Consider any strategy $\langle \bar{e}^k \rangle$ that doesn't do anything in i and doesn't require any resource inputs; it's trivial to see that this is feasible. Since A is indifferent to i, we do some algebraic substitution of utility values to see that the conjugate plan has strictly higher utility than $\langle \bar{e}^k \rangle$ does. This means that $\langle \bar{e}^k \rangle$ is suboptimal.

Theorem 3: Defection Threshold

Even if U is not indifferent about i, we can prove Theorem 2 can still hold if the utility A gains by enacting the aforementioned conjugate strategy is greater than the maximal amount U cares about i (formally expressed by $\max_{s,s' \in S_i} |U_i(s) - U_i(s')|$).

Proof. The structure is similar to that of Theorem 2; just use the new maximum caring differential in the algebraic substitution.

> We interpret Theorem 3 as a partial confirmation of Omohundro's thesis in the following sense. If there are actions in the real world that produce more resources than they consume, and the resources gained by taking those actions allow agents the freedom to take various other actions, then we can justifiably call these actions "convergent instrumental goals." Most agents will have a strong incentive to pursue these goals, and an agent will refrain from doing so only if it has a utility function over the relevant region that strongly disincentivizes those actions.

The Bit Universe

The authors introduce a toy model and use the freshly-proven theorems to illustrate how A takes non-null actions in our precious S_h (both when it is indifferent to h and when it is not). This isn't good; the vast majority of utility-maximizing agents will not steer us towards futures we find desirable. If you're interested, I recommend reading this section for yourself, even if you aren't very comfortable with math.

Our Universe

> The path that our model shows is untenable is the path of designing powerful agents intended to autonomously have large effects on the world, maximizing goals that do not capture all the complexities of human values. If such systems are built, we cannot expect them to cooperate with or ignore humans, by default.

We have much work to do. The risks are enormous and the challenges "impossible", but we have time on the clock. AI safety research is primarily talent-constrained. If you've been sitting on the sidelines, wondering whether you're good enough to learn the material - well, I can't make any promises. But if you feel the burning desire to *do something*, to put forth some extraordinary effort, to become stronger - I invite you to contact me so we can work through the material together.

Questions and Errata

- Page 4, left column, last line: why is that $\cup P^t$ - shouldn't we take the union of the outputs and whatever resources *weren't* used at time t?
- Page 8, right column, second full paragraph, last line: should be "we have two options available *to* us".

[1] By the axiom of substitution, $\text{Feasible}(\langle P^k \rangle) = \{\langle \bar{a}^k \rangle : \text{isFeasible}(\langle P^k \rangle, \langle \bar{a}^k \rangle)\}$.

A Proper Scoring Rule for Confidence Intervals

You probably already know that you can incentivise honest reporting of probabilities using a proper scoring rule like log score, but did you know that you can also incentivize honest reporting of confidence intervals?

To incentize reporting of a 90% confidence interval, take the score $-S - 20 \cdot D$, where S is the size of your confidence interval, and D is the distance between the true value and the interval. D is 0 whenever the true value is in the interval.

This incentivizes not only giving an interval that has the true value 90% of the time, but also distributes the remaining 10% equally between overestimates and underestimates.

To keep the lower bound of the interval important, I recommend measuring S and D in log space. So if the true value is T and the interval is (L, U), then S is $\log(\frac{U}{L})$ and D is $\log(\frac{T}{U})$ for underestimates and $\log(\frac{L}{T})$ for overestimates. Of course, you need questions with positive answers to do this.

To do a P% confidence interval, take the score $-S - \frac{2000}{100-P} \cdot D$.

This can be used to make training calibration, using something like Wits and Wagers cards more fun. I also think it could be turned into app, if one could get a large list of questions with numerical values.

Mythic Mode

Follow-up to: The Intelligent Social Web

Related to: Fake Frameworks

Yesterday I described a framework for viewing culture as a kind of distributed intelligence, and ourselves as nodes in this distributed network.

Today I'd like to share a way of using this framework intentionally that doesn't require Looking. My main intent here is concreteness: I'd like to illustrate what an application of accounting for the Omega-web can look like. But I also hope this is something some of y'all can benefit from.

I'll warn up front: this is playing with epistemic fire. I think the skill of clearly labeling when you're entering and leaving a fake framework is especially important here for retaining epistemic integrity. If you aren't sure how to do that, or if the prospect of needing to unnerves you too much, then it might be right for you not to try using this at least for now.

Scott Alexander created a fascinating impact through his essay Meditations on Moloch. A few excerpts:

> *What's always impressed me about this poem is its conception of civilization as an individual entity. You can almost see him, with his fingers of armies and his skyscraper-window eyes.*
>
> […]
>
> *The Universe is a dark and foreboding place, suspended between alien deities. Cthulhu, Gnon, Moloch, call them what you will.*
>
> *Somewhere in this darkness is another god. He has also had many names. In the Kushiel books, his name was Elua. He is the god of flowers and free love and all soft and fragile things. Of art and science and philosophy and love. Of niceness, community, and civilization. He is a god of humans.*
>
> *The other gods sit on their dark thrones and think "Ha ha, a god who doesn't even control any hell-monsters or command his worshippers to become killing machines. What a weakling! This is going to be so easy!"*
>
> *But somehow Elua is still here. No one knows exactly how. And the gods who oppose Him tend to find Themselves meeting with a surprising number of unfortunate accidents.*
>
> *There are many gods, but this one is ours.*

There was basically nothing in that essay that was conceptually new, at least in the social circles I'm in around CFAR. But this essay still had a *huge* cultural impact. Suddenly it became real how to literally see the demon-god so many of us are fighting, and now we know its name: *Moloch*. This *mattered* to the web. Now we can

actually *feel* a sense in which we're battling eldritch horrors in an epic war determining humanity's future.

I suggest that the reason this is impactful comes from something I mentioned in my last post: the social web encodes its sense of meaning, roles, and expectations in the structure of story. Facts can inform culture, but story *guides* it. Scott's main contribution via that essay, I claim, was in his transposition of large chunks of the fight against existential risk into the key of myth.

From this mythic mode, within the sandbox, we can see a sense in which classical gods are real. We can see the footprints of Ares, the ferocious warrior, in the bomb-carved craters of areas torn up by civil war. Or Apollo's light in the lively intelligent discourse between academic colleagues who are sincerely curious about the truth. Or the joint partnership of Dionysus and Hephaestus that breathes life into the crazy builder revelry that is Burning Man. To the extent that something like these archetypes are known and recognizable to each of us in our story-like intuitions, these gods can be seen as distributed subroutines within the web of Omega.

The programmer Eric S. Raymond describes this beautifully in an old essay Dancing With the Gods. I really recommend reading the whole thing; he's quite lucid about it. Here's a relevant snippet:

> *All the Gods are alive. They are not supernatural; rather, they are our inmost natures. They power our dreams and our art and our personalities. Theurgy and ritual can make them stronger, more accessible to the shaman. They can be evoked in a human being to teach, heal, inspire, or harm. Occasionally they manifest in spontaneous theophanies; the result may be religious conversion, creative inspiration, charisma, or madness.*

Mythic mode is a way of looking at the world through a story-like lens. When you enter mythic mode, you recognize that you're a character in Omega's story, as is everyone else. And because you're very likely familiar with a wide range of story types, you can probably look around and see who has been given which kind of plot hook, and to what kind of tale.

Why is any of this relevant?

Well, recall from my previous post that there's a basic puzzle: if you don't like the script you're enacting, you won't get very far just trying to defy it, because by default your effort to defy it will just play into your role.

But… we *do* have *stories* of people being able to transition roles in a pretty deep sense. They often (but not always) follow the arc of the hero's journey, wherein the hero must enter into the unknown and face trials and eventually die to who they were, transforming into something new so as to complete the journey and return victorious but different. We tell these (or similar) stories again and again, with lots of variation… but some things (like the types of heroes) tend to vary a lot less than others. This gives us some clues about where the web has room to let people shift their lived scripts, and what the constraints are.

So, if you can identify a well-known story type that fits the *transition* you want and also starts from a place pretty close to where you are, and you have enough slack to lean into that role, then the web might conspire to help you play out that script.

The thing is, you can't just sit outside your role and figure out what to do. That isn't what it feels like to live the epic you're examining; that's playing the role of someone who is (among other things) analyzing the story they *think* they're in.

Instead, if you want to use this approach, you have to learn how to *experience story from the inside*. That's essentially what mythic mode is.

I like how Eric Raymond expressed this part too (again from Dancing With the Gods):

> *If my language is too "religious" for you, feel free to transpose it all into the key of psychology. Speak of archetypes and semi-independent complexes. Feel free to hypothesize that I've merely learned how to enter some non-ordinary mental states that change my body language, disable a few mental censors, and have me putting out signals that other people interpret in terms of certain material in their own unconscious minds.*
>
> *Fine. You've explained it. Correctly, even. But you can't do it!*
>
> *And as long as you stick with the sterile denotative language of psychology, and the logical mode of the waking mind, you won't be able to --- because you can't reach and program the unconscious mind that way. It takes music, symbolism, sex, hypnosis, wine and strange drugs, firelight and chanting, ritual and magic. Super-stimuli that reach past the conscious mind and neocortex, in and back to the primate and mammal and reptile brains curled up inside.*

I think it's important afterwards to be able to *leave* mythic mode, and leave the "insights" gleaned within the sandbox, and give your more normal way of interpreting the world a chance to look at what happened. In particular, mythic mode tends to highlight seemingly meaningful coincidences, but at least some of those are likely to be confirmation bias, which is helpful to remember once you're outside the sandbox.

But I think it's also critical *not to do this while in mythic mode*. It just gets in the way. You in fact don't know ahead of time which synchronicities are confirmation bias and which are you syncing up with the larger computational network, and it's *too slow* in practice to figure it out in real time, and the effort of trying tends to shove you into a role type that won't let you walk the path of a hero's journey you weren't already on anyway. You are in fact donning some epistemic risk whenever you use mythic mode — which is why I think it's important to sandbox it properly if you're going to bother.

I'd like to illustrate the use of mythic mode with a personal example to help clarify what it can look like.

Right after my kenshō, I tried to find a teacher in Rinzai Zen, since that's the tradition I'm familiar with that treats kenshō as an initiation point after which deeper instruction becomes possible. This turned out to be tricky: Sōtō Zen (with its emphasis on gradual development and its downplaying of the relevance of kenshō) is *so* much more popular that Rinzai dojos basically don't exist anywhere near where I live, at least that I could tell.

This felt weird. I'd reached kenshō via a previous arc of using mythic mode, and finding a Rinzai teacher felt like the natural next step, but I was getting stuck. This "plot has led me to a dead end" feeling has become a signal to me to switch into mythic mode to try reinterpreting the blockade.

From mythic mode, I considered what kind of character I was, including the implicit genre-savviness I was using. When I imagined wearing the role of a zen disciple and walking the path to becoming a zen master, I noticed that it almost but didn't quite fit my sense of my path, like I'd be being a little dishonest to who I am. I focused on the "not quite right" feeling, and what came up was my love of physicality and athletics… and *martial arts*. And there's *totally* an archetype for someone who walks the path of enlightenment via martial arts: the Eastern warrior-monk. That felt right. From mythic mode, then, it seemed promising for me to see how to walk *that* path.

Some Googling suggested to me that the origin of this archetype was the Shaolin Monastery. It seems that their spiritual practice was Chan Buddhism, from which we get all the schools of zen. This closely matched the "almost but not quite right" feeling I'd gotten earlier. In mythic mode, this is the kind of thing I've learned to take as evidence that I'm going in a mythically supported direction. (From outside mythic mode, it's really not that surprising that I'd find something like this that I could interpret as meaningful… but since at this point I hadn't solved the original problem, I wasn't going to worry too much about that just yet.)

After a sequence of mythic exploration and omens, it seemed clear to me that I needed to visit New York City. I was actually ready to hop on a plane the day after we'd finished with a CFAR workshop… but a bunch of projects showed up as important for me to deal with over the following week. So I booked plane tickets for a week later.

When I arrived, it turned out that the Shaolin monk who teaches there was arriving back from a weeks-long trip from Argentina *that day*.

This is a kind of thing I've come to expect from mythic mode. I could have used murphyjitsu to hopefully notice that maybe the monk wouldn't be there and then called to check, and then carefully timed my trip to coincide with when he's there. But from inside mythic mode, that wouldn't have mattered: either it would just work out (like it did); or it was fated within the script that it *wouldn't* work out, in which case some problem I *didn't* anticipate would appear anyway (e.g., I might have just failed to think of the monk possibly traveling). My landing *the same day* he returned, as a result of my just *happening* to need to wait a week… is the kind of coincidence one just gets used to after a while of operating mythically.

(And of course, this is quite possibly just confirmation bias. And that's important to notice. But like I said earlier, one tends to get results from mythic mode if one isn't too worried about that *while in the mode*. And also, we don't know it *is* confirmation bias either: what people notice, and when, is subject to the distributed computation of the social web, which means that some seeming coincidences are probably orchestrated. E.g., maybe some part of me noticed a sidebar on their website mentioning when the monk would be traveling, but I didn't consciously register it, instead feeling like those little projects I had to take care of were important enough to have me wait a week.)

The whole trip in New York felt epic. I gained a *lot*. Most of what I gained requires more backstory to explain, so for the sake of brevity I'll skip describing the bulk of it. I did learn an intense movement meditation sequence I've been using almost every morning for months now — which, interestingly, I don't need to struggle to get myself to do. I just get up and do it, easily. It's not a matter of personal discipline; it's just so right-fitting for me that it happens naturally.

Looking back, from *outside* mythic mode, I can see how this amounted to me doing a costly self-signal to stick to some kind of meditation and exercise program. That fits

with most of the insights and opportunities I experienced along the way. And… I can also see how it wouldn't have worked if I'd done it thinking "I'm going to spend a bunch of time and money on a costly signal to myself." I step outside mythic mode and keep its "insights" contained within the sandbox as a matter of keeping my epistemology clean. But even from here, I can see how valuable that toolkit is to me as a method of shaping my behavior and, sometimes, getting myself to update.

I've seen the rationality community use mythic mode a lot — but almost exclusively for intuition pumps and, occasionally, spicing up events. And even then I've seen a fair amount of push-back. My guess, from extrapolating the outcries I've heard against it, is that a fair number of folk find it epistemically frightening. And that makes sense: if you don't know how to sandbox, or if you don't trust that sandboxing can reliably work, then this probably looks like a crazy risk to take.

From where I'm standing, though, the choice was already made when you were born. We're *already* embedded in culture and subject to its influences. And much of that is culture reaching into our emotions and deepest thoughts and nudging us to behave in certain ways. None of us are immune; if it were otherwise, *there would be no reason for caution*.

The fact that there is a *type of person* who is attracted to Less Wrong, and that this type *gathers* and forms a *community*, but that the vast majority of that community is not and has not been involved in the same task-oriented project… suggests that the forces that shape the rationality community are implicit and subtle, and probably very similar to the ones that shape other communities.

So from what I can tell, if you don't know how to sandbox this stuff, and you don't know how to Look, then your epistemology is *already* screwed. It just might not be in-character for you to notice it in this way.*

With all that said, I'm not at all invested in folk here using mythic mode more than they already do. I wanted this here to illustrate an example application of accounting for the real-world Omega. I'll also want to call on the framework later to offer my own intuition pumps in future posts: it's a really helpful context for conveying maps that point at otherwise-hard-to-talk-about phenomenology.

Beyond that, if you want to avoid using mythic mode, I don't object.

I even welcome attempts to argue that *no one* should use mythic mode. Just be warned that you're likely to find that effort a particular flavor of frustrating.

---

*: In case I need this later, this is an MD5 hash: 24e07349c9134ff91d77a6a38cf23183

Two types of mathematician

This is an expansion of a linkdump I made a while ago with examples of mathematicians splitting other mathematicians into two groups, which may be of wider interest in the context of the recent elephant/rider discussion. (Though probably not *especially* wide interest, so I'm posting this to my personal page.)

The two clusters vary a bit, but there's some pattern to what goes in each - it tends to be roughly ‘algebra/problem-solving/analysis/logic/step-by-step/precision/explicit’ vs. 'geometry/theorising/synthesis/intuition/all-at-once/hand-waving/implicit’.

(Edit to add: 'analysis' in the first cluster is meant to be analysis as opposed to 'synthesis' in the second cluster, i.e. 'breaking down' as opposed to 'building up'. It's not referring to the mathematical subject of analysis, which is hard to place!)

These seem to have a family resemblance to the S2/S1 division, but there's a lot lumped under each one that could helpfully be split out, which is where some of the confusion in the comments to the elephant/rider post is probably coming in. (I haven't read *The Elephant in the Brain* yet, but from the sound of it that is using something of a different distinction again, which is also adding to the confusion). Sarah Constantin and Owen Shen have both split out some of these distinctions in a more useful way.

I wanted to chuck these into the discussion because: a) it's a pet topic of mine that I'll happily shoehorn into anything; b) it shows that a similar split has been present in mathematical folk wisdom for at least a century; c) these are all really good essays by some of the most impressive mathematicians and physicists of the 20th century, and are well worth reading on their own account.

- The earliest one I know (and one of the best) is Poincare’s ‘Intuition and Logic in Mathematics’ from 1905, which starts:

> “It is impossible to study the works of the great mathematicians, or even those of the lesser, without noticing and distinguishing two opposite tendencies, or rather two entirely different kinds of minds. The one sort are above all preoccupied with logic; to read their works, one is tempted to believe they have advanced only step by step, after the manner of a Vauban who pushes on his trenches against the place besieged, leaving nothing to chance.
>
> The other sort are guided by intuition and at the first stroke make quick but sometimes precarious conquests, like bold cavalrymen of the advance guard.”

- Felix Klein's 'Elementary Mathematics from an Advanced Standpoint' in 1908 has 'Plan A' ('the formal theory of equations') and 'Plan B' ('a fusion of the perception of number with that of space'). He also separates out 'ordered formal calculation' into a Plan C.
- Gian-Carlo Rota made a division into ‘problem solvers and theorizers’ (in ‘Indiscrete Thoughts’, excerpt here).
- Timothy Gowers makes a very similar division in his ‘Two Cultures of Mathematics’ (discussion and link to pdf here).
- Vladimir Arnold's ‘On Teaching Mathematics’ is an incredibly entertaining rant from a partisan of the geometry/intuition side - it's over-the-top but was 100% what I needed to read when I first found it.
- Michael Atiyah makes the distinction in ‘What is Geometry?’:

> Broadly speaking I want to suggest that geometry is that part of mathematics in which visual thought is dominant whereas algebra is that part in which sequential thought is dominant. This dichotomy is perhaps better conveyed by the words “insight” versus “rigour” and both play an essential role in real mathematical problems.

There’s also his famous quote:

> Algebra is the offer made by the devil to the mathematician. The devil says: `I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine.’

- Grothendieck was seriously weird, and may not fit well to either category, but I love this quote from *Récoltes et semailles* too much to not include it:

> Since then I’ve had the chance in the world of mathematics that bid me welcome, to meet quite a number of people, both among my “elders” and among young people in my general age group who were more brilliant, much more ‘gifted’ than I was. I admired the facility with which they picked up, as if at play, new ideas, juggling them as if familiar with them from the cradle – while for myself I felt clumsy, even oafish, wandering painfully up an arduous track, like a dumb ox faced with an amorphous mountain of things I had to learn (so I was assured), things I felt incapable of understanding the essentials or following through to the end. Indeed, there was little about me that identified the kind of bright student who wins at prestigious competitions or assimilates almost by sleight of hand, the most forbidding subjects.
>
> In fact, most of these comrades who I gauged to be more brilliant than I have gone on to become distinguished mathematicians. Still from the perspective or thirty or thirty five years, I can state that their imprint upon the mathematics of our time has not been very profound. They’ve done all things, often beautiful things, in a context that was already set out before them, which they had no inclination to disturb. Without being aware of it, they’ve remained prisoners of those invisible and despotic circles which delimit the universe of a certain milieu in a given era. To have broken these bounds they would have to rediscover in themselves that capability which was their birthright, as it was mine: The capacity to be alone.

- Freeman Dyson calls his groups ‘Birds and Frogs’ (this one’s more physics-focussed).
- This may be too much partisanship from me for the geometry/implicit cluster, but I think the Mark Kac 'magician' quote is also connected to this:

> There are two kinds of geniuses: the ‘ordinary’ and the ‘magicians.’ an ordinary genius is a fellow whom you and I would be just as good as, if we were only many times better. There is no mystery as to how his mind works. Once we understand what they’ve done, we feel certain that we, too, could have done it. It is different with the magicians... Feynman is a magician of the highest caliber.

The algebra/explicit cluster is more 'public' in some sense, in that its main product is a chain of step-by-step formal reasoning that can be written down and is fairly communicable between people. (This is probably also the main reason that formal education loves it.) The geometry/implicit cluster relies on lots of pieces of hard-to-

transfer intuition, and these tend to stay 'stuck in people's heads' even if they write a legitimising chain of reasoning down, so it can look like 'magic' on the outside.

- Finally, I think something similar is at the heart of William Thurston's debate with Jaffe and Quinn over the necessity of rigour in mathematics – see Thurston's 'On proof and progress in mathematics'.

Edit to add: Seo Sanghyeon contributed the following example by email, from Weinberg's *Dreams of a Final Theory*:

> Theoretical physicists in their most successful work tend to play one of two roles: they are either sages or magicians... It is possible to teach general relativity today by following pretty much the same line of reasoning that Einstein used when he finally wrote up his work in 1915. Then there are magician-physicists, who do not seem to be reasoning at all but who jump over all intermediate steps to a new insight about nature. The authors of physics textbook are usually compelled to redo the work of the magicians so they seem like sages; otherwise no reader would understand the physics.

Two Coordination Styles

In game theory, assumptions of rationality imply that any "solution" of a game must be an equilibrium.* However, most games have many equilibria, and realistic agents don't always know which equilibrium they are in. Certain equilibrium strategies, such as *tit-for-tat* in iterated prisoner's dilemma, can also be seen in this broader context as *coordination strategies*: adopting them teaches others to adopt them, because you punish anyone playing some other strategy. In a narrow sense, these strategies solve both the game itself and the equilibrium selection problem. (Technically, such strategies are the evolutionarily stable ones.)

I want to make an informal point about two very different ways this can work out in real life: coordination strategies in which it feels like everyone is fighting to pull the system in different directions but it all cancels out, vs situations where it feels like the coordination strategy is your friend because it saves everyone effort. I believe the second case exists, but it is rather puzzling in terms of the existing literature.

**Different rationality assumptions give different equilibrium concepts; Nash equilibria are the most popular. Correlated equilibria are the second most popular and somewhat more relevant to the discussion here, but I won't get into enough technical details for it to matter.*

This post made possible by discussions with Steve Rayhawk, Harmanas Chopra, Jennifer RM, Anna Salamon, and Andrew Critch. Added some edits proposed by Elo.

Schelling Negotiations

Schelling discussed agents solving the equilibrium selection problem by choosing points which other agents are most likely to choose based on prominence, and the term *Schelling point* was coined to describe such likely equilibria. The classic examples revolve around agents who cannot communicate with one another (highlighting the need for guesswork about each other's behavior), but adding the ability to communicate does not eliminate the equilibrium-selection problem; our community tends to use the term 'Schelling *fence*' to refer to the analogous concept when open negotiation is involved -- though my impression is that economic literature uses Schelling point for this case as well.

In *A Positive Account of Property Rights*, David Friedman** explains the emergence of property rights through the negotiation of Schelling fences. Negotiators want as many resources as possible, but also want to minimize the costs of conflict. If there were nothing special about any one patch of land, then both sides could always demand a little more land -- there's no good stopping-point for the bargaining. However, natural divisions in the land such as rivers can serve as Schelling fences. Once such a solution has been proposed, neither side wants to demand a little more for themselves, because breaking the Schelling fence opens up the door for the other person to do the same.

Even if the territory is blank, Schelling fences can be made with abstract reasoning: a half-half split, for example, is simpler than other options.

The coordination problem *does* re-assert itself at a higher level: gerrymandering conceptual categories. Is the river or the rocky ridge the more natural dividing line? What really counts as the "border" of the rocky ridge, exactly? Participants will tend to engage in motivated cognition to justify whichever potential Schelling fence is a better fit for them.

The side-taking hypothesis of morality, discussed in *The Side-Taking Hypothesis for Moral Judgement* by Peter DeScioli and *A Solution to the Mysteries of Morality* by DeScioli and Kurzban, gives a similar account of where our moral intuitions come from. This time, rather than just thinking about two people negotiating a conflict about property rights, we think about the bystanders. Bystanders may get caught up in a conflict, as the two contestants call on their allies for support. People may have their own interests in the outcome, but they also prefer to end up on the winning side rather than the losing side. So, everyone is trying to predict which side others will take, in order to choose sides. This creates an equilibrium-selection situation, so simple rules about right and wrong can dominate complicated social calculations. (As before, complex social considerations come back due to the possibility of gerrymandering the concepts which make up the Schelling points.)

Ziz's blog has good discussions of social order and justice in a Schelling-type framework, and coins the term *Schelling reach* to quantify a population's coordination power (how complex can one's reasoning be and still settle on the same Schelling point as others?). We can also understand language as an equilibrium-selection game: when you try to say something, you have to balance the various plausible interpretations which your audience might place on the various options of language at your disposal. People will gerrymander word meanings to make their preferred arguments more compelling. Ziz discusses consequences of this in DRM'd Ontology.

Willpower as Self-Coordination

Now let's relate this to arguments you have within yourself, via Ainslie's explanation of willpower in *Breakdown of Will*. Ainslie gives evidence that humans have systematically different preferences at different points in time. Moods and drives make different things desirable. Easy pleasure gets more tempting as it gets nearer. You set an alarm at night thinking it would be good to get up early and get more done, yet you prefer to shut it off and sleep in in the morning.

Ainslie suggests that we view this as a negotiation between instances of you across time. He calls these "interests" to avoid constantly sounding like he's talking about multiple personality disorder. Ainslie's definition of *willpower* is successful coordination between these interests via bright-line rules. Each interest knows that if it breaks the rules, it breaks your ability to coordinate with yourself; although each interest has somewhat different goals, the threat of global *inability to coordinate* is great enough to balance against almost any temptation.

This is why willpower feels sort of like a top-down imposition: some interests are blackmailing other interests with the threat of global discoordination, to make yourself do something which you don't want to do right now but which fits with your concept of "what you want to do".

One problem with this is that you have to use simple rules. Why? It's a lot like the Schelling negotiations discussed earlier. The interests have to coordinate on an equilibrium, which means it must be simple enough that there aren't plausible

alternatives to bargain for. (Although, interests may try to bend even the simplest rules. There's a special Schelling-art to calculating excuses. "This is a special occasion, I can break my diet just this once!")

Anna Salamon has described these systems of rules as *ego structures* rather than willpower (private communication). I really like thinking of it this way, and had written a post draft about it, but it wasn't written very well so I may or may not post it.

A second problem is that in order to threaten yourself with a global coordination failure, you have to be willing to follow through. This isn't such a problem with humans, because habit-building works this way already: it isn't very plausible that you'll be able to build healthy habits in the future if you keep breaking them in the present. However, greater threats provide greater incentives. This makes some people engage in more directly self-punishing behavior if they don't live up to their own standards, such as mentally beating themselves up and feeling awful about things, depriving themselves of other pleasures, etc.

Two Coordination Styles

Ziz strongly advises against Ainslie's willpower strategy, calling it self-blackmail. Nate Soares seems to do the same in the replacing guilt sequence. Most of Ziz's blog is about an alternative technique called fusion. (I don't recommend just reading that link; read Ziz from the beginning. The posts before the fusion posts are prerequisites.) Nate similarly spends most of his blog explaining how to do better. Both of them use self-coordination strategies which have *aspects* of Ainslie's approach: they view themselves as made up of sub-agents, and explicitly think about the coordination of those sub-agents. However, the flavor is much more like building up trust between sub-agents rather than blackmail. Other people in the bay area rationality community also seem to advocate similar approaches, particularly Andrew Critch (in in-person conversations). It seems like the people who do this have more Getting Stuff Done power. But how could this work? Schelling's framework seems rather compelling. Is there some way around it?

So, I've finally got to the point I promised to make at the beginning: it seems like there are two different sorts of coordination strategies. I don't have any better terminology lined up, so as per Schelling-nature, I'll borrow Ziz's: treaties vs fusion.

- ***Treaties:*** Ainslie's, Friedman's, and DeScioli & Kurzban's coordination mechanism involves bright-line rules, guilt vs innocence, retributional justice, and inefficient compromises (though Friedman argues in *Law's Order* that approximate efficiency is often achieved). Norms are established by historical precedence. Gerrymandering and casuistry abound.
- ***Fusion:*** Nate's, Critch's, and Ziz's coordination style involves totality of circumstances (the legal term for the opposite of bright-line rules -- careful consideration of the circumstance). Their parts cooperate more fully, rather than spending a lot of the time fighting against each other. Decisions look closer to utility calculations rather than historical precedence. Action comes from intrinsic motivation rather than self-imposed structure.

How is this possible? What are they doing differently? If I buy Ainslie's psychological model even approximately, this seems rather difficult to explain.

Here are several ideas:

1. A legalistic system with punitive justice is more relevant to a society made up of a large number of agents who come and go, and so are only able to build trust to a limited degree. All the parts of a human live in the one head, so it is possible, though by no means universal, that a human can come to trust themselves.
2. People play tit-for-tat with other people in lots of ways, tracking how much they owe and are owed in dollars and favors, social status and personal trust. However, while this accounting facilitates an efficient and fair society, tracking every parties' balance can itself be costly. With certain people, there's a kind of mutual wink that happens, and you implicitly or explicitly agree to drop the whole accounting structure and just optimize the joint utility function. When you do that, new things become possible; you no longer have to make sure that each person gets their due in the end, which is actually better for both of you in expected value (potentially much better). This is described well, in my estimation, by true friendship is being counterfactually hugged by vampires on Compass Rose.
3. Ainslie's model recommends that we make a very important conceptual shift. Don't just view actions as isolated. See them as instances of the class of similar actions (as in my instance of class post). Why? If you see your action as a part of a pattern, you'll consider its consequences on the coordination game. Only then can self-blackmail work. This is similar to CDT vs EDT in twin prisoner's dilemma: CDT thinks as if it can change its action in isolation without changing the probability of its copy cooperating, whereas EDT thinks of its action as correlated to its other self. In the same way, you have to recognize that your actions now are correlated with your actions later. *However*, perhaps Ainslie is missing another level of this kind of thinking: you know that if you gerrymander a rule to serve your narrow interests right now, your selves across time will similarly gerrymander. If you stop gerrymandering, you can apply totality-of-circumstances rather than bright-line thinking; it is more easily manipulated, but you have agreed with all your selves to stop trying to manipulate things. The very realization that you can achieve fusion rather than treaty, and that fusion is much better than treaty, is enough to incentivise all your parts to coordinate in this way.

***David Friedman is an expert in the economic analysis of law, and draws striking relationships between what rules are economically efficient, what's intuitively just, and what's used in practice. His book* Law's Order *contains more in this direction.*

The Monthly Newsletter as Thinking Tool

At the start of 2014 I accidentally started writing a monthly newsletter for a handful of close friends, and I've mostly kept it up since then. It was originally supposed to be kind of commitment device for myself, a way of holding myself accountable to my 2014 goals by commiting to sharing my progress with those friends at the end of each month. It continued to serve that function for the next four years, but also became something much more useful than that.

The easiest, very high ROI step is to simply open up a new Evernote document on the first of each month and title it "February 2018 Update" or whatever. Then just try to keep that tab open. You'll find stuff to put in there. What I end up writing generally falls into a small number of categories:

- Links to things that I think my friend would be interested in, with a bit of discussion of why I think it's interesting. These also serve as future reference material for me.
- Updates on my progress toward some goal or another, usually written in a style meant to be at least readably entertaining. This lets me look back over the years and see exactly what I was doing and when.
- Discovery of some new thing that obsesses me briefly that prompts me to write 10,000 words of evangelism about it (e.g. meditation, hypnosis, trigger point therapy, Alexander technique, Ghokale method posture, jiu-jitsu, longevity supplements, some new AI architecture, Mr. Rogers, EverQuest as an exemplar of Fun Theory, the ketogenic diet, at least four different exercise regimens) which serves as very useful reference material, which I tend to frequently refer back to when the topic ends up being something that I make part of my life.
- At one point I started writing a story for my friends and sending it to them in installements, which gradually turned into a 50,000 word book over the course of a year.

You don't even have to share it with anybody, I suppose, but I suspect you'll find it much more motivating to actually write in it if you do intend to share it. I created a private Blogger blog in which all these monthly newsletters reside. You could probably do the same thing with a public blog.

That might prompt you to remark, "Congratulations, genius, you've rediscovered blogging." Perhaps that's fair. But I think it does help that I know I'm only writing for a small handful of close friends. For one thing, it gives me a specific audience to write for, which helps focus my thinking. For another, I feel less inhibited, because I know that no matter what I write, I won't end up having to defend myself from a troll.

And the final and most important distinction between a monthly newsletter and a blog -- and, I think, the place where all the value of this practice comes in -- is the time-locked nature of it. You have a month to jot down thoughts, then at the end of the month you have to "finish" those thoughts. Process all the garbage you may have dumped into your Evernote document into something that your friends will actually want to read all the way through. This accomplishes a lot of things.

- It's a cure for the goodism that leads you to have fifteen unfinished drafts of blog posts that never see the light of day. You've got a month. Wrap it up.
- You're forced to organize your random notes and copy-pastes and unlabeled links into something readable enough for others that it ends up passing as powerful reference material for your future self.
- Your thinking on a given topic will "advance" in a way that it otherwise wouldn't. You're forced to finish and polish the stubs of thought processes that you may have thrown in. (A lot of this finishing and processing is happening subconsciously throughout the month. If you know you're going to have to share it at the end of the month, your brain will give you something worth sharing. Without that pressure, your thoughts just tend to continue on, going in circles for years without ever resulting in anything useful to even yourself. Or at least, mine do.)

And if all those specific benefits don't convince you, I suppose I'll add that I feel that the last four years would've been quite impoverished if I didn't have this habit. Or meta-habit, because it's become such an important part of my life that the newsletters are how I keep myself honest regarding my other important habits and goals. Plus I have four years of records of my own life and dense notes on a variety of topics that interest me, that I otherwise wouldn't have.

Beware Social Coping Strategies

Worlds

The world is a huge, complex, scary place.

Winter used to freeze us dead, and today lives are still destroyed by natural disasters like tsunami and forest fires. Plagues used to destroy our communities, and people still die from disease every day. We want more power over our surroundings — more intelligent computers, better space rockets — but construction is difficult and often dangerous. As a child, navigating the world directly is also terrifying: sharp objects, distressingly loud sounds, heights, electric shocks, etc. It takes an incredible amount of thought to merely *understand* and *predict* the world, let alone to avoid being hurt by it.

The social world is even worse.

Every person is a world in themselves. Every person has their own model of the world, plus their own thoughts and dreams and desires and morality. These may not always be consistent with your own, and that can cause conflicts. Their preferences are not always consistent with the preferences of other people, and conflicts between them can affect you. Individuals can also have conflicts inside themselves, making them unpredictable and confusing.

And it's not just individuals you have to navigate. When people get together, they form groups and institutions. You are born, and find it's not just sharp objects you have to watch out for, but violating codes that were made up by people you've never met. You find yourself in complicated systems with their own logic and rules: government, school, a family structure, religious or political groups. These systems are additional worlds to learn — and they share the frustrating property of individuals that they are not entirely consistent.

Navigating all these worlds, and learning how to follow or evade their rules without getting hurt, is tough.

Being Thwarted

As if that weren't bad enough, the social world has a special property that the natural world doesn't have: adaptive problems. That is, problems which adapt when you try to solve them[1].

In the natural and abstract world, problems are broadly static. If you move a rock out of your way, it stays there. When you realise 1+1=2, it doesn't change its mind and become 3. But if you try to solve a problem in the social world, people are constantly changing the problem landscape. They can intentionally thwart you — using their own creativity to adapt the problem and make it harder for you to solve.

It's like you solve the problem of stubbing your toe on rocks by looking ahead more, and then they notice this and *jump out* to meet your foot. (So then you make sure to avoid rocky paths... only to find the rocks have formed a committee to ban walking around flat ground.)

How might this look (in lieu of bureaucratic rocks)?

Your parent puts cookies on the top shelf. You solve this using your genius understanding of the natural world: objects can be moved and climbed on — so you push a chair and triumphantly retrieve your cookie. But your parents are like magic cookie-quest homing missiles, which *somehow* know when you're about to succeed in your quest and then put endless barriers in your way: put them higher, take away your broom, lock them in a cabinet, eat them first.

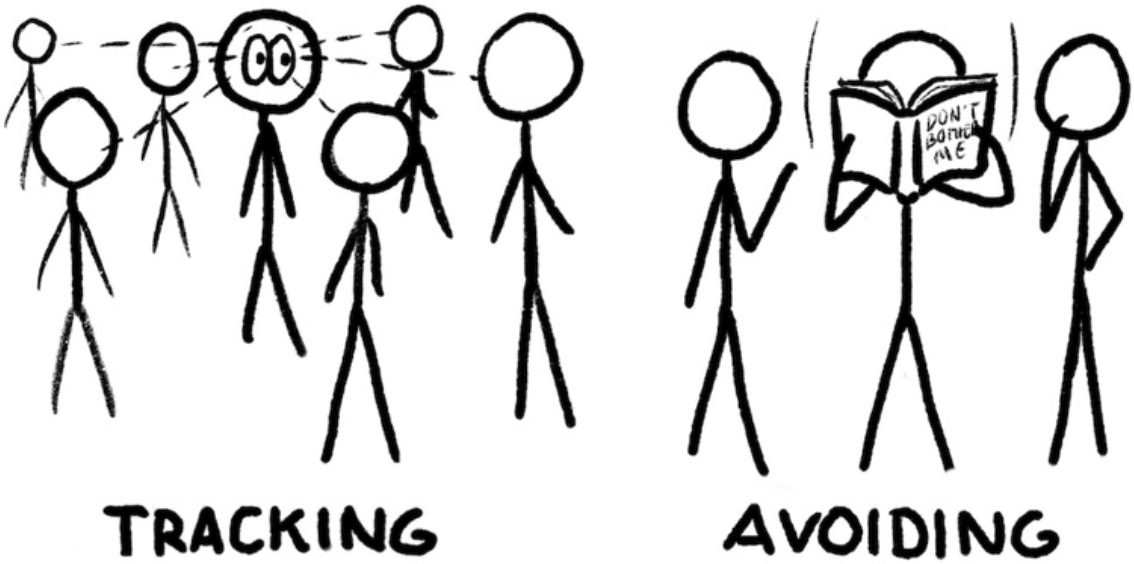
Then you discover that you can get the cookies if you solve the problem in a different way: doing things that please the parents! Pleasing parents is a difficult problem. They seem to respond well to being adorable — but also to being impressive? Hmm, these seem to push in different directions. They act impressed when you solve this block puzzle, but they smile when you give them a cuddle when they're making that frowny-face. But sometimes they don't smile and instead tell you to go away. Confusing. Okay, so either try ignoring them (maybe they will leave you alone if you don't respond?), or try to understand this parent-moods thing...

Two Coping Strategies

When we go through life, we develop coping strategies for dealing with the reality of all these worlds that are other human beings.

Two common coping strategies for dealing with the social world are:

1. **Be hyper-attentive / track it.** Put effort (consciously or not) into understanding and reading people — how they think, how they respond, what they want.
2. **Avoid it.** Block out the social world, and focus on the analytic world instead.



Both of these coping strategies come from an over-sensitivity to the presence of other people's wishes — an inability to hold one's own wants and other people's in mind at the same time (without feeling conflict/pressure).

Both of these coping strategies come with large costs. But in return, they act to make the world a little more predictable. If you can't get something nice, you can at least get something predictable. When the choice is between 'fairly bad and unpredictable' and 'even worse, but at least predictable', there is incentive to choose the latter: in those situations, one can make progress (you can build your life around a sucky-but-predictable world); but in an unpredictable world, progress is difficult — being thwarted takes a constant stream of creativity to overcome.

These aren't the only two strategies. Some people, for instance, find ways of being focused on the analytic world while also feeling at ease socially. By *coping* strategy, I mean to indicate it's not a full solution — it's a bandage, a way of managing. (I discuss solutions in the last section.) It's also common to use a mix of strategies, or use different strategies in different situations. People may not fall strictly into one category or another.

1. Hyper-Attentive Tracking

People who adopt the first strategy cope by being hyper-aware of and sensitive to the people around them. They are preoccupied by the wants of others; and one way this plays out is that they try to keep everyone happy, even at the expense of their own wants.

This coping strategy tends to cause problems with anxiety (tracking people is hard — so much info! — and people, even individuals, often want contradictory things, and it feels bad when there is conflict). Other common problems with this coping strategy include: discomfort when alone; lack of strong sense of self (own needs taking less priority than the needs of

others); lacking interests that can be done without people; difficulty regulating emotions, and highly influenced by the emotions of other people.

It can feel like being at the mercy of others' wishes. Counter-intuitively, the way this coping strategy works is by controlling them: understand someone's wishes enough to be what they want, so that they do what you want them to do (such as accept you, or stop being in conflict with you).

If you're hurt by someone, you naturally try to control them: make them stop hurting you. Or if not make them stop, at least make it predictable. For example, some forms of defiant naughtiness in schools involve deliberately provoking a teacher in order to get punishment. When punishment is predictable, you can just do what you always do, instead of having to creatively come up with new ways to fend off threats all the time.

Coping strategies often start in adverse circumstances, like the above, but then continue to be used even in neutral or friendly circumstances. People find themselves being manipulative or inauthentic and don't know why. It then takes effort and practice to re-learn authentic ways of interacting.

2. Avoiding

People who adopt the second strategy cope by chronically averting their attention from the wants of other people. Their coping strategy is to *limit the amount of information coming in*.

For instance, some people close or avert their eyes from their conversation partner when speaking. In most conversations, there is a lot of information being exchanged even from the silent party: whether they're following along, whether they want to interrupt, how they're feeling about what you're saying, etc. If you have trouble with knowing how to incorporate this information while not losing your own thoughts, the simplest solution is to limit it.

They may also try to limit the amount of information going *out*, too. Monotone speech, neutral expression, static body language — don't move a muscle, or else people will be able to read your mind! (Not realising that you ain't fooling no one — people can read your mind anyway.)

This strategy cripples connections with other people, which require a lot of back-and-forth of information. People who cope this way usually know they can have trouble in groups and with making friends — or at very least they find social situations and people stressful to deal with.

Inner World

Someone with the avoidant coping strategy might be content with a small, chosen group of friends. They're happy to only make connections with certain kinds of people, and generally focus on their own pursuits over playing 'social games'.

If this were the only problem with this limiting-information coping mechanism, it wouldn't be a big deal — you'd just have a less-social lifestyle. Yeah, you might get lonely sometimes, or not get high in certain career ladders, but it's not a bad life.

But it's worse than that. The problem is bigger:

The different sides of you also have that information leak.

We have lots of different parts of ourselves. Whenever we're feeling conflicted, that can be thought of as two sides having a disagreement. [2]

Internal disagreements come up a lot between the conscious or explicit part of our mind vs the subconscious or inexplicit part of our mind. Thoughts vs feelings. System 2 vs System 1.

In reality, these shouldn't be warring sides. A well-adjusted person takes seriously *all* sides of their mind, and doesn't privilege one side or the other. If there's a disagreement, simply declaring one side is right is an epistemologically authoritarian move: *you can't know* which side is right before resolving the conflict between them.

The coping strategies that sabotage the information flow from other people *also* sabotage the information flow inside a single mind.

This is why 'social' and 'feelings' are often lumped together (contrasted with 'intellectual'). Social interaction uses the same mental logic as navigating disagreeing thoughts and feelings inside one mind.

Why do some people struggle with motivation to work on their goals? Because they're in a highly conflicted state. Many people get so good at severing the connection between their emotions and thoughts that they can't even tell what what they want or what they're conflicted about. It's just this vague sense of deadness.

Sabotaging this information flow can feel like internal discomfort, or suppression, or feeling stuck, or stress, or boredom, or lacking in motivation or focus. Taking feelings and forbidden thoughts seriously can feel terrifying, so we shut them down.

Your relationship with other people is a macrocosm of your relationship with yourself. [3]

You're just even better at shutting off that information when it's internal.

Secure Solution

It's possible to get through life without resorting to either of these coping strategies. Someone can be a happy hermit — not blocking people out, but not focusing on them either. Or someone could be quite social, but not feel anguish or pressure when people want things from them, and instead take the "different strokes for different folks" attitude: we can agree to disagree, you're okay, I'm okay, it's all okay.

When you're a child, you really are totally reliant on adults to provide for your needs. You may not have the option to 'agree to disagree'. You may experience school as being forced in a room with other people you have to interact with, by threat of punishment. If you're unlucky, you won't find a nice way of navigating this, and you'll form a coping strategy to make it bearable.

But when you're an adult, you are independent. You have choice to decline interactions you find unpleasant. You don't need everyone you know to like you to have a functioning life. There are still people and institutions to navigate, but they aren't out to get you. They won't thwart your cookie quests. You are free.

---

[1] CFAR uses the phrase 'adaptive problems' to mean a slightly different but related concept:

"*A problem whose solution contains steps or methods that are unknown or uncertain, often requiring experimentation, novel strategies, or entirely new ways of thinking.*"

Here, I'm emphasising problems that adapt to become harder when you try to solve them. Problems that adapt responsively.

[2] CFAR uses this concept for their Internal Double Crux method of resolving internal conflict.

[3] Here, I've been talking about how the avoidant coping strategy has consequences for how you deal with different parts of your own mind.

But the logic of the attentive coping strategy could also have consequences: Under this coping strategy, if you have internal conflicts, you may be inclined to control the different parts of your own mind. This may manifest as hyper-vigilance over your own thoughts.

I think this kind of controlling attentiveness may manifest in other ways, but I haven't come up with many examples yet. If anyone has any thoughts about this, leave them in the comments!

Beware arguments from possibility

For any claim X, exactly one of these is true:

1) X is possible - compatible with the evidence so far.

2) X is impossible - there's a contradiction between X and the evidence so far.

That means asserting the possibility of something is harder than you think. For example, if Bob says: "Epiphenomenalism is possible, therefore <far reaching conclusions>"

Wait, Bob, did you just say it's *impossible* to find a contradiction between epiphenomenalism and anything else you know? That's an awfully strong claim! You got any evidence?

Bob pulls back: "I meant only that X sounds plausible to me." But that's a fact about Bob's limits of reasoning, it doesn't support the far-reaching conclusions anymore. For that you need to justify (1) over (2), not just assert it.

Factorio, Accelerando, Empathizing with Empires and Moderate Takeoffs

I started planning this post before Cousin It's post on a similar subject. This is a bit of a more poetic take on moderate, peaceful AI takeoffs.

Spoilers for the game Factorio, and the book Accelerando. They are both quite good. But, if you're not going to get around to play/reading them for awhile, I'd just go ahead and read this thing (I think the game and story are still good if you go in knowing some plot elements).

i. Factorio

Factorio is a computer game about automation.

It begins with you crash landing on a planet. Your goal is to go home. To go home, you need to build a rocket. To build a rocket powerful enough to get back to your home solar system, you will need advanced metallurgy, combustion engines, electronics, etc. To get those things, you'll need to bootstrap yourself from the stone age to the nuclear age.

To do this *all by yourself,* you must automate as much of the work as you can.

To do this efficiently, you'll need to build stripmines, powerplants, etc. (And, later, automatic tools to build stripmines and powerplants).

One wrinkle you may run into is that there are indigenous creatures on the planet.

They look like weird creepy bugs. It is left ambiguous how sentient the natives are, and how they should factor into your moral calculus. But regardless, it becomes clear that the more you pollute, the more annoyed they will be, and they will begin to attack your base.

If you're like me, this might make you feel bad.

During my last playthrough, I tried hard not to kill things I didn't have to, and pollute as minimally as possible. I built defenses in case the aliens attacked, but when I ran out of iron, I looked for new mineral deposits that didn't have nearby native colonies. I bootstrapped my way to solar power as quickly as possible, replacing my smog-belching furnaces with electric ones.

I needed oil, though.

And the only oil fields I could find were right in the middle of an alien colony.

I stared at the oil field for a few minutes, thinking about how *convenient* it would be if that alien colony wasn't there. I stayed true to my principles. "I'll find another way", I said. And eventually, at much time cost, I found another oil field.

But around this time, I realized that one of my iron mines was *near* some native encampments. And those natives started attacking me on a regular basis. I built

defenses, but they started attacking harder.

Turns out, just because someone doesn't *literally* live in a place doesn't mean they're happy with you moving into their territory. The attacks grew more frequent.

Eventually I discovered the alien encampment was... pretty small. It would not be that difficult for me to destroy it. And, holy hell, would it be *so* much easier if that encampment didn't exist. There's even a sympathetic narrative I could paint for myself, where so many creatures were dying every day as they went to attack my base, that it was in fact merciful to just quickly put down the colony.

I didn't do that. (Instead, I actually got distracted and died). But this gave me a weird felt sense, perhaps skill, of *empathizing with the British Empire.* (Or, most industrial empires, modern or ancient).

Like, I was trying really hard not to be a jerk. I was just trying to go home. And it *still* was difficult not to just move in and take stuff when I wanted. And although this was a video game, I think in real life it might have been if anything *harder*, since I'd be risking not just losing the game but losing my life or livelihood of people I cared about.

So when I imagine industrial empires that *weren't* raised by hippy-ish parents who believe colonialism and pollution were bad... well, what realistically would you expect to happen when they interface with less powerful cultures?

ii. Accelerando

Accelerando is a book about a fairly moderate takeoff of AGI.

Each chapter takes place 10 years after the previous one. There's a couple decade transition from "complex systems of narrow AIs start being relevant", "the first uploads and human/machine interfaces", to "true AGI is a major player in the Earth and solar system."

The outcome here... reasonably good, as things go. The various posthuman actors adopt a "leave Earth alone" policy - there's plenty of other atoms in the solar system. They start building modular chunks of a dyson sphere, using Mercury and other hard-surface planets as resources. (A conceit of the book is that gas giants are harder to work with, so Jupiter et al remain more or less in their present form)

The modular dyson sphere is solar powered, and it's advantageous to move your computronium as close as possible to the sun. Agents running on hardware closer to the sun get to think faster, which lets them outcompete those further away.

There are biological humans who don't do any kind of uploading or neural interfacing. There are biological-esque humans who use various augmentations but don't focus all their attention on competing on the fastest timescales with the most advanced posthumans.

The posthumans eventually disassemble all the terrestrial planets and asteroids. What's left are the gas giants (hard to dissassemble) and Earth.

Eventually (where by "eventually" I mean, "in a couple decades"), they go to great lengths to transport the surface of Earth in such a way that the denizens get to retain

something like their ancestral home, but the core of the Earth can be used to build more computronium.

And then, in another decade or two (many generations from the perspectives of posthumans running close to the Sun's heat), our posthuman offspring take another look at this last chunk of atoms...

...and sort of shrug apologetically and wring their metaphorical hands and then consume the last atoms in the solar system.

(By this point, old-school humans have seen the writing on the wall and departed the solar system)

iii. Moderate Takeoffs

I periodically converse with people who argue something like: "moderate takeoff of AGI is most likely, and that there'll be time to figure out what to do about it (in particular if humans get to be augmenting themselves or using increasingly powerful tools to improve their strategizing)."

And... this just doesn't seem that comforting to me. In the most optimistic possible worlds I imagine (where we don't get alignment exactly right, but a moderate takeoff makes it easier to handle), human level AI takes several decades, the people who don't upload are outcompeted, and the final hangwringing posthuman shrug takes... maybe a few centuries max?

And maybe this isn't the *worst* outcome. Maybe the result is still some kind of sentient posthuman society, engaged in creativity and various positive experiences that I'd endorse, which then goes on to colonize the universe. And it's sad that humans and non-committed transhumans got outcompeted but at least there's still *some* kind of light in the universe.

But still, this doesn't seem like an outcome I'm *enthusiastic* about. It's at least not something I'd want to happen by default without reflecting upon whether we could do better. Even if you're expecting a moderate takeoff, it still seems really important to get things right on the first try.

The abruptness of nuclear weapons

Nuclear weapons seem like the marquee example of rapid technological change after crossing a critical threshold.

Looking at the numbers, it seems to me like:

- During WWII, and probably for several years after the war, the cost / TNT equivalent for manufacturing nuclear weapons was comparable to the cost of conventional explosives, (AI impacts estimates a manufacturing cost of $25M/each)
- Amortizing out the cost of the Manhattan project, dropping all nuclear weapons produced in WWII would be cost-competitive with traditional firebombing (which this thesis estimates at 5k GBP (=$10k?) / death, vs. ~100k deaths per nuclear weapon) and by 1950, when stockpiles had gown to >100 weapons, was an order of magnitude cheaper. (Nuclear weapons are much easier to deliver, and at that point the development cost was comparable to manufacturing cost).

Separately, it seems like a 4 year lead in nuclear weapons would represent a decisive strategic advantage, which is much shorter than any other technology. My best guess is that a 2 year lead wouldn't do it, but I'd love to hear an assessment of the situation from someone who understands the relevant history/technology better than I do.

So my understanding is: it takes about 4 years to make nuclear weapons and another 4 years for them to substantially overtake conventional explosives (against a 20 year doubling time for the broader economy). Having a 4 year lead corresponds to a decisive strategic advantage.

Does that understanding seem roughly right? What's most wrong or suspect? I don't expect want to do a detailed investigation since this is pretty tangential to my interests, but the example is in the back of my mind slightly influencing my views about AI, and so I'd like it to be roughly accurate or tagged as inaccurate. Likely errors: (a) you can get a decisive strategic advantage with a smaller lead, (b) cost-effectiveness improved more rapidly after the war than I'm imagining, or (c) those numbers are totally wrong for one reason or another.

I think the arguments for a nuclear discontinuity are really strong, much stronger than any other technology. Physics fundamentally has a discrete list of kinds of potential energy, which have different characteristic densities, with a huge gap between chemical and nuclear energy densities. And the dynamics of war are quite sensitive to energy density (nuclear power doesn't seem to have been a major discontinuity). And the dynamics of nuclear chain reactions predictably make it hard for nuclear weapons to be "worse" in any way other than being more expensive (you can't really make them cheaper by making them weaker or less reliable). So the continuous progress narrative isn't making a strong prediction about this case.

(Of course, progress in nuclear weapons involves large-scale manufacturing. Today the economy grows at roughly the same rate as in 1945, but information technology can change much more rapidly.)

An alternative way to browse LessWrong 2.0

This is something I've been tinkering with for a while, but I think it's now complete enough to be generally useful. It's an alternative frontend for LessWrong 2.0, using the GraphQL API.

Features:

- Fast, even on low-end computers and phones
- Quickly jump to new comments in a thread with the "." and "," keys
- Archive view makes it easy to browse the best posts of years past
- Always shows every comment in a thread, no need to "load more"
- Log in and post using your existing username and password, or create a new account
- Simple markdown editor
- Typography enhancements
- Switch between fixed-width and fluid layouts and several different themes
- Easily view a comment's ancestors without scrolling by hovering over the left edge of a comment tree

Thanks to Said Achmiz for designing the themes and writing much of the frontend JavaScript.

Give it a try: https://www.greaterwrong.com

Rationalist Lent

It's that time of year again. Pick something to give up for 40 days as an experiment / comfort zone expansion. Post about it here. Good luck and have fun!

Edit: For anyone who wants it, here are some prompts for brainstorming things to try giving up (set a 5-minute timer for the brainstorm):

1. What takes up too much of my time?
2. What takes up too much of my attention?
3. What in my life is most out to get me?
4. What superstimuli am I most attracted to?
5. What in my life feels the most like pica?

And also for anyone who wants it, here are some concrete suggestions for things to give up, in no particular order:

1. Video games
2. Smartphones
3. Facebook, Reddit, etc.
4. Sugar
5. Caffeine

Pain, fear, sex, and higher order preferences

Suppose a footballer with a broken leg is steeling themselves to kick a penalty despite the pain. Suppose a statistics lecturer is about to take a plane to a conference, and is trying to overcome their fear of flying. Or suppose a manager is meeting a client, and is trying to overcome their sexual attraction and stay professional.

All three situations can be seen as a conflict between higher order rationality and lower order urges, but they are actually different. In every case, there is an *instrumental* reason to overcome the "urge", but the general attitude towards that urge is different.

For the footballer, they want to overcome their current pain avoidance instincts, but they don't want to get rid of those instincts entirely. They generally want to avoid pain, but, for the moment, they have something more important to do - win a game - and would want to put their pain aversion preference aside for the moment.

In contrast the statistician with a fear of flying wouldn't mind having that fear expunged for ever. They know that normal flying involves no large risks, so a fear of flying is wholly irrational.

Thus we see pain aversion as an endorsed preference; we wouldn't want to get rid of it. Fear of flying is an unendorsed preference: we would toss it out if we could.

What of the third example? Well, that depends on what the manager feels about their own sexual urges. They may be more or less endorsed, depending on various factors of the manager's values and social circumstances.

Higher order preferences?

For fear of flying, we can fit that fear into a simple narrative of lower order preferences being overruled - or not - by higher order rationality.

But it's not clear that pain aversion (or sexual desire) can be fit into the same narrative. Since we don't want to override our pain aversion, it might be fitting to see pain aversion as higher order preference itself - or at least "not overriding pain aversion" might be a higher order preference, while pain aversion is a lower order one.

One can complicate the picture further by considering more edge cases, but the simplest interpretation is that the simple "higher order versus lower order" does not fit pain aversion or (often) sexual desire.

On Building Theories of History

This is an excerpt from the draft of my upcoming book on great founder theory. It was originally published on SamoBurja.com. You can access the original here.

Why was Barack Obama elected president in 2008? Was it because he ran a smart and successful campaign? Was it because new social media sites allowed young people to get interested in politics? Or was it because American culture was generally shifting away from George W. Bush's brand of conservatism?

If you read the news articles published on November 4th, 2008, you'll notice something interesting: journalists explain this historic event in many different ways. Some journalists attribute the campaign's success primarily to the individual leading the campaign. Others focus more on the influence of new technologies on campaigning. Still others explain it in terms of a general cultural or political shift.

These explanations are revealing—not necessarily of what actually landed President Obama in office, but rather of how each individual journalist conceives of the way things happen in the world. Through their explanations for the outcome of the election, we can glean a bit of their implicit theories of history.

Concept & importance

A **theory of history** is an explanation of how things generally happen in the world, both in the past and in the future. If, for example, you subscribe to the great man theory of history, then you might explain events by looking at the influential individuals who shaped them. If you subscribe to a technological determinist theory, on the other hand, you might explain events in terms of the technologies that allowed them to occur. Or, you might subscribe to a social determinist theory, explaining both influential individuals and new technologies as the makings of greater societal forces.

Someone operating under the great man paradigm might explain Obama's election as a product of his and his staff's exacting efforts in the day-to-day of campaign work. A believer in technological determinism might attribute the win to the unprecedented use of social media, which mobilized previously uninterested voters. Someone adhering to a social determinist view might draw a straight line from the Civil Rights era to the election of Barack Obama, pointing out the inexorable cultural shift towards empowering African-Americans.

Everyone has a theory of history, in that everyone has an explanation of why the world is how it is and an understanding of how the world changes and has changed. Everyone has to: without an understanding of how the world works, no matter how faulty, implicit, or subconscious, we would be prohibited from acting in what we believe is the right way to achieve our goals, whether big or small. Few people could tell you plainly that they are social or technological determinists, or adherents of great man theory. But everyone, if asked, can give reasons why some event or another happened, and whether, or why, it might happen again in the future.

We don't just explain things with our theories of history. We act on them. If you believe that individuals have the power to significantly shape history, for example, you might be more inclined to make things happen yourself. If you believe that technology drives historical change, you might specifically try to invent new technologies. If, on the other hand, you believe that the fate of the world has already been decided, or if you believe that history is inevitably heading in a certain direction, you may be less inclined to take a stand. After all, if it's going to happen, then it's going to happen.

Therefore, whether we're trying to change the world in a major way or just live our lives in society in the best way possible, it's vital that we come to understand the true theory of history. We need the true theory of history in order to take the right actions in the world, and we need to accurately predict the results of our actions. If we have an incorrect theory of history, we run the risk of producing unknown and possibly catastrophic consequences, for ourselves or others.

It's important here to note the distinction between the *true* theory of history, and the "true" theory of history that we're aiming for. The *true* theory of history will be unmanageably complex, because the number of factors that actually influence what happens in the world is incalculably large. Because of its complexity, the *true* theory of history will be difficult, if not impossible, to use to explain what's going on in the world. In aiming for the "true" theory of history, we are assuming the power law: we are assuming that there will be a small number of factors that have disproportionately large effects on the world, or that can explain the existence of other factors. We are aiming for a theory that generally explains how things happen in the world. Going forward, we will drop the quotation marks and stipulate that the true theory of history is the theory that takes into account the core causes contributing to the world as it exists, making it comprehensible and usable to us mere mortals.

No one has it

No one in the world has figured out the true theory of history. If they did, we'd know: not only would they be extremely, visibly, powerful, but they would be active in many domains—politics, religion, culture, technology—reshaping society step by step, or taking seemingly prescient advantage of trends, with many successes and few false starts. There are historical examples of incredible individuals, such as the Indian Emperor Ashoka the Great, and organizations, such as the Catholic Church, whose repeated success across multiple domains is difficult to explain without them understanding at least fragments of a true theory of history.

There are many reasons why no one has figured out the full true theory of history, some psychological and some practical.

There are at least three psychological reasons for why most people are deterred from finding the true theory of history. The first is that the vast majority of people only have

an implicit, or subconscious, theory of history. In other words, most people do not even have the *concept* of a theory of history. The problem with relying on your implicit theory of history is that it's wrong, without a doubt. The world is complex, and your theory of history has to explain how everything in the world works. Without explicitly trying to improve your theory of history, there is no hope: there will be countless things that you have not had the time or perspective to take into account. Improving your theory of history implicitly is not systematic enough to work.

The second reason why no one has managed to achieve the true theory of history is that many people endorse one theory of history while unknowingly acting on another. For example, some people explicitly endorse the technological determinist view of history even as they implicitly act on the great man paradigm: believing that it will require the work of remarkable individuals to create the technology that will save the world, for example, instead of believing that the inevitable, impersonal progress of technology will do so.

There can be many belief-based reasons for why people fall into this trap, but on a more basic level, people simply don't have a good sense of what their implicit theories of history are, or know how to explicate them, which means they cannot reliably align their intellectual and emotional beliefs. To some extent, acting on your implicit theory of history while operating under a different explicit theory is fine — after all, your implicit theory will, for a while, be more nuanced than your explicit one. What is problematic is to act unconsciously on one theory of history and proclaim another; this makes it very difficult to improve your implicit theory of history, which you act on.

The third reason is that people tend to switch between theories of history in an unprincipled way, which prevents them from noticing theory-threatening anomalies. If they can't notice and explain seeming anomalies in their theory of history, then they can't improve their theory. If someone largely adheres to the great man paradigm, for example, but resolves any contradictions by falling back on the technological determinist view, then they've prevented themselves from justifying their understanding of the great man theory, or realizing that their justification is inadequate or incorrect. Theory-threatening anomalies have to be resolved, not rationalized.

These are just a few of the psychological barriers that prevent people from making progress towards the true theory of history. But there's a simpler, more practical problem: the world is complex. In order to understand it, you need the right methodology, and you need a huge amount of properly processed data.

Read more from Samo Burja here.

Science like a chef

Alice: Hey honey, I made pasta with tomato sauce!

Bob: Great, let's eat!

Bob: Mmmmm, that's fantastic! It's even better than last time. It's got a sweeter, deeper flavor, which I like.

Alice: Thanks. Last time I only sautéed onions and garlic before adding the tomato puree, but this time I added carrots to the mix for some extra sweetness.

Bob: You know I love your cooking, but I always feel a twinge of skepticism whenever you try to explain why things taste the way they do. Yes, you added carrots to the sauté. But is that the only thing you did differently?

Alice: No. I added some butter along with the olive oil to give it some smoothness. Ummm. I threw some red wine in to the sauce. We didn't have any basil leaves today so I just left that out of the recipe.

Bob: So how do you know that the *carrots* are what made it taste sweeter? How do you know that it wasn't the butter, the red wine, or the absence of basil leaves?

Alice: Well, the red wine is acidic, so that wouldn't make it taste sweet. I think the butter helped make it taste richer and smoother. The basil leaves *add* flavor, so leaving them out probably subtracted from the deepness, rather than adding to it.

Bob: Sure, that all sounds plausible. But how can you *know* any of this?

Alice: It's just common sense. I know that butter is rich. I know that red wine is acidic. I know that basil is flavorful. So I know that adding them will have the corresponding effects on the sauce.

Bob: But isn't it possible that things work differently when you *combine* them? Sure, red wine is acidic by itself, but isn't it possible that when combined with sautéed onions, garlic and carrots, and olive oil and butter, and fried tomato paste, and whatever spices you used, and cooked. Isn't it possible that after all of that, some chemical reactions occur that cause the red wine to be sweet instead of acidic?

Alice: I don't think so. I think the sweetness came from the carrots. But yeah, I suppose it is *possible*. So what do you propose I do?

Bob: Cook like a scientist. Only change one variable at a time. Take note of the effect. So in this case, if you hypothesize that adding carrots to the sauté will add sweetness to the sauce, *just* add carrots to the sauté. *Don't* add red wine, or anything else.

Alice: Ok, let's suppose I did do that, and I found that carrots did in fact add sweetness. What if I want to add red wine next time. Isn't it possible that carrots add sweetness to the basic sauce, but when combined with red wine, carrots react with the red wine and actually add, I don't know, bitterness to the sauce?

Bob: Yes! You're cooking like a scientist now!

Alice: So I'm supposed to then do *another* experiment? Make the sauce with red wine, make the sauce with red wine and carrots, and see what the effect of carrots are in that particular scenario?

Bob: Absolutely.

Alice: And then what if I want to use ghee for my fat instead of olive oil?

Bob: More experiments.

Alice: Don't you think that so much experimenting is impractical?

Bob: Maybe. But it's the only way to know the truth. The laws of scientific inference don't care how easy they are to follow. They're *laws*.

Alice: Maybe. Have you ever heard of bayesian inference?

Bob: Uh yeah, I think so. I think they had a brief section on it in one of my undergrad stats classes. I remember something about it helping solve a game show problem. But we don't use it much in our work at the lab.

Alice: Well the reason I bring it up is because you seem to view inference as a binary thing. Either you conclude that carrots added sweetness to the sauce, or you don't. That's not how I, or other bayesians see things. To me, everything is a spectrum. Of course I can't say that I'm *100%* sure that the carrots were the cause of the resulting sweeter flavor, but I will say that I'm about 90% sure. Yes, the fact that I added red wine, butter and removed basil introduce the possibility of confounding, but I feel like I have a pretty good handle on these ingredients given my experience cooking, and my judgement is that it's unlikely that they were the cause of the extra sweetness.

Bob: *Sigh*. I mean, I guess you're right. But that's just a very unscientific way of doing things.

Alice: And why is that a bad thing?

Bob: Because with that approach, you can never know for *certain* what the truth is. Like you said, you're only 90% sure that the carrots caused the sweetness, not 100% sure.

Alice: Well I'm ok with that. In a perfect world, I'd do things your way and only manipulate one variable at a time, but that's just not practical in the kitchen. Like you said, every time I change one thing, I'd have to redo the experiment with the carrots. Given how many variables there are, I'd have to do hundreds, maybe even thousands of experiments.

Bob: Again, proper experimentation is the only way to actually know things with certainty.

Alice: Yes, proper experimentation would allow me to boost my confidence from 90% to 100%*. But the cost of this boost would be hundreds of experiments. Thousands of hours of my time. I don't care enough to do that. The return on investment is totally not worth it. When the return on investment isn't worth it, sometimes it makes sense to do science like a chef.

Bob: Well when I cook, I like to do science like a scientist.

Alice: And *that's* why your tomato sauce stinks.

Lessons from the Cold War on Information Hazards: Why Internal Communication is Critical

Due to their tremendous power, nuclear weapons were a subject of intense secrecy and taboo in the US government following WW2. After their first uses, President Truman came to consider atomic weapons a terror weapon of last resort and generally avoided discussion of their use.[1] As a consequence, top-level decision makers in his administration (himself included), outside the Atomic Energy Commision (AEC) remained unaware of the US nuclear stockpile size as it grew,[2] and relatively few means of achieving arms control were considered at a high level, given Stalin's resistance to opening up the Soviet Union to inspections.[3] Frustrated by failed negotiations and geopolitical expansion by the communists, US atomic weapons were made ready for use again in 1947,[4] and no arms control talks were attempted again beyond a propaganda level for the next ten years.[5]

The consequences of a lack of internal communication, however, did not end with Truman. Somewhat insulated from Eisenhower's stance against nuclear first use,[6] Curtis LeMay's plans with Strategic Air Command grew increasingly toward it.[7] Growing from a 1000 weapon air force to an 18,000 weapon triad[8] with up to potentially 27,000 megatons of explosive power by 1960[9] the arsenal went far beyond the necessity needed for Eisenhower's massive retaliation strategy. If the warheads were equally sized and distributed, an airburst attack with such a number of weapons would have been capable of scorching more than half of the Soviet wilderness or shattering windows over more than 90% of the USSR, China, and North Korea.[10] While one might argue that such a degree of force may have been rational for accounting for reliability, or retaining significant retaliatory capability after a first strike by the Soviet Union, LeMay's plans presumed using nearly all the weapons in a preemptive strike[11] as US bombers were vulnerable on open runways. Generally speaking, nuclear war planners of the 50s were uncoordinated, didn't account for many of the known effects of nuclear weapons, and were given no war objectives to plan toward from which limits could be derived.[12] Though Eisenhower learned that SAC's plans were excessive even for counterforce, Eisenhower didn't commission any study on the lasting environmental effects of such a massive use of force.[13]

With secrecy and taboo around nuclear weapons, each party involved in planning their use and policy ended up with a great deal of autonomy, where they could make mistakes which would be impossible in a higher feedback environment. With a lack of idea spread between different parts of government, the Strategic Air Command could pursue plans and procurements that were not necessarily efficient or aligned with US political interests. Though analysts at the RAND Corporation had determined counterforce strategies were problematic,[14] as SAC didn't share it's intelligence abilities, it could feel overly confident dismissing such analysis.[15]

Since the US's open society likely had greater vulnerability to Soviet spies than vice versa, it makes sense that the US was concerned about leaks which could give the USSR an advantage or create public pressure to force the president into taking non-strategic actions. In general, compartmentalizing information is a great way to reduce

the risk of spies leaking a decisive amount of information, but cutting information flow at the highest levels of government directly harms the government's ability to make decisions on the basis of the classified information it possesses. If one can't trust high-level planners enough to give them access to the information needed to make rational plans, such people should not have been planners in the first place. If too many people are required to make good plans securely, then compartmentalization should take place based on accomplishing smaller objectives and creating models from which plans can be derived. Instead, compartmentalization happened by military service branch and agency, resulting in duplication of planning effort, and too many inexperienced people being involved without criticism.

To characterize the disunity of intelligence and communication during the Cold War some in the defense community use this joke:[16]

> **US Air Force:** *"The Russians are here!"*
>
> **Defense Intelligence Agency:** *"The Russians are not here yet, but they are coming."*
>
> **CIA:** *"The Russians are trying, but they won't make it."*
>
> **Intelligence and Research (INR, in the Department of State):** *"The Russians? They aren't even trying."*

For states and research organizations to not fall prey to these sorts of mistakes in areas situations with risks of harmful information spread (information hazards), there are a few principles that seem like they can be derived from this.

1: It is extremely important for people with good ideas to press them within their secure communities in order to improve decisions. Ideas kept secret and immune from criticism are likely to be pretty bad or ill-conceived, but if there is risk from public engagement (eg. by spreading knowledge of a bioweapon that may have become easier to manufacture) then such conversations should still be had within secure communities to better flesh out the real risks and create strategies for mitigating them.

2: The bounds of secrecy should be well defined. Generally stifling conversation shuts down both useful and harmful information flow, so making good delineations can result in a net reduction in risk. Secrecy can be abused by groups to gain information advantages against competing interest groups, to increase one's social status, or to maintain corruption. For this reason, cultures of openness and transparency can sometimes develop a net advantage against more secretive ones since they better align the incentives of those involved.

Footnotes:

1. Rosenberg, D. A. (1983). The Origins of Overkill: Nuclear Weapons and American Strategy, 1945-1960. International Security, 7(4), 11.
2. ibid., 11
3. McGeorge Bundy, Danger and Survival, (New York: Random House, 1988), 130

4. ibid., 339
5. ibid., 130
6. ibid., 252
7. ibid., 322
8. ibid., 319
9. ibid., 320
10. Calculation made using land area estimations from google maps and Alex Wellerstein's nukemap app using 1.5 Megaton airburst weapons. Note that as all weapons would not be uniform, the actual maximum area damage would be lower.
11. Bundy 322
12. Moore, John H. (14 February 1957). "Letter from Captain John H. Morse, Special Assistant to the chairman, Atomic Energy Commission, to Lewis Strauss, Chairman, Atomic Energy Commission". In Burr, William. "It Is Certain There Will be Many Firestorms": New Evidence on the Origins of Overkill (PDF). Electronic Briefing Book No. 108 (Report). George Washington University National Security Archive. Dwight D. Eisenhower Library, Records of Special Assistant for National Security Affairs, NSC Series, Briefing Notes Subseries, box 17, Target Systems (1957–1961).
13. Bundy 324-325
14. Andrew David May, 'The RAND Corporation and the Dynamics of American Strategic Thought, 1946–1962', PhD Dissertation, Emory Univ. 1998, 235, 291.
15. Austin Long & Brendan Rittenhouse Green (2015) Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy, Journal of Strategic Studies, 38:1-2, 44, DOI: 10.1080/01402390.2014.958150
16. Johnson, L. K. (2008). Glimpses into the Gems of American Intelligence: The President's Daily Brief and the National Intelligence Estimate. Intelligence and National Security, 23(3), 333-370. doi:10.1080/02684520802121257

The Logic of Science: 2.2

This is a linkpost for http://pulsarcoffee.com/posts/los2-2/

This is the first technical post on my new blog, where I plan to continue writing about what I learn from self-study of mathematics. I'm currently reading E.T. Jaynes' *Probability Theory: The Logic of Science*, and Christopher Bishop's *Pattern Recognition and Machine Learning*. While I understand many LessWrong2.0 readers are far above my level in maths, maybe there are some who would benefit from and enjoy conversation about the sort of things I'm learning.

Open-Source Monasticism

[I originally wrote most of the strings of text below for the online *Art & Monasticism Symposium* in 2012 through **Transpositions,** a collaborative effort of students associated with the *Institute for Theology, Imagination, and the Arts* at the University of St Andrews. What follows has been edited since it was first published, for tone as well as content. I used to write a lot about this topic on my blog (RIP *otherhood.org)*. So I may be porting some old posts to this new version of LW.]

Monasticicismsms

Assertion: Monasticism is a recurring pattern in the world's religions. Looking at the things that different monasticisms have in common can teach us how (and how not) to live in communities of shared purpose.

[This word, *monasticism,* basically means "the way of life of people living in monasteries." And to further tighten down the jargon filter, by *monastery* I mean the class of objects that includes cloisters, abbeys, nunneries, convents, etc. and related things like ashrams and kibbutzes, but not necessarily all intentional communities. My favorite word in English for a person who lives in a monastery, even temporarily, is *monk* (which I hereby and for all time declare to be a gender-neutral word for nuns and male-nuns).]

Monasticism := Monks + Monasteries

While there is an ENORMOUS amount of variation and localized ornamentation, this monastic tendency–for serious practitioners of a religious tradition to band together to benefit from living in a disciplined community–shows up all over the place, e.g. in most strains of Christianity and Buddhism.

Why would this be the case? Clearly there are economic efficiencies. It’s easier to survive as a community, in which specialization becomes possible, than alone; some monks are good at holding the big picture, while some monks are better at painting icons, and others are virtuosos at cooking nourishing meals, or farming, or carpentry, or web design.

There are efficiencies that have to do with spiritual practice as well. It is much easier to hold a daily routine of work and contemplation when you do so with a group of like-minded brothers or sisters.

And when singing or chanting, of course, it is only possible to make harmonies together (unless you're throat-singing).

Stepping Meta

From this perspective, monasticism starts to look less like something spiritual and more like a type of technology, an arrangement of hardware and software that has arisen in various places at various times to meet certain universal needs—for stable community and material support, as well as the sublime. We need the nectar of

transcendence as much as we need more tangible kinds of nourishment, and monastic life tries to provide both.

Like technology, monasticism appears to evolve over time as groups' needs change, and as social, cultural, political, and ecological climates shift from season to season.

Some monastic hardware & software is better suited for north Africa in the 6th century, while another kind fits medieval Japan after the arrival of Buddhism, and a dozen others are appropriate to the array of niches in today's postmodern religious world.

Applied, Secularized Monasticism

I've been working with a NON-religious group of misfits collectively known as the Art Monastery Project over the past decade to (begin to) take the source code of this software, the blueprints of this hardware, and apply them to art-making and the creative process (as well as to the pursuit of wisdom and compassion, whatever those words mean).

Our goal, not totally different from that of some other monastic orders, is to cultivate personal awakening and cultural transformation through art, community, and contemplation.

Thankfully for us, most monastic software is open source, freely available to anyone with eyes to see.

And this means that anyone could be doing something analogous but for purposes other than or in addition to art-making [like maybe, oh I dunno... rationality?], using the same source code, for well or ill. If you are, we should be talking.

So how are we doing it? Read on if you're curious about what monastic technologies we have tried using to run our art projects.

Our monastic hardware/software stack

Community**,** both as a social arrangement and as the physical space that houses it, is one of the biggest pieces of monastic technology.

Yet a monastery, whether Benedictine or Kagyu, is not the same as an intentional community. As a *structure in space*, a monastery has intentionality designed into every square foot. As a *structure in time*, a routine that carries practitioners toward greater balance, compassion and wakefulness, a monastery must be flexible and free enough to adapt to changing circumstances and to give monks time to be quiet and still, yet rigid enough to keep monks active and productive.

Thus the ideal monastic structure in space-time is informed by the history of monastic architecture, as well as the routines that have guided monks for hundreds of years.

Discipline is another piece of technology we apply to art-making. At certain times of the year we get up at the same time, work together to sustain ourselves as well as to make art (be it music, dance, theater, painting, sculpture, performance art, code,

poetry, prose, synchronized swimming, etc.…), and engage in contemplative activities like meditation, chant, and reading together throughout the day.

For the past 10 years, we have run public "artmonk retreats" and private practice periods, as well as laboratories for connecting silent contemplation and creativity, inspiration and expression. We have experimented with asceticism and renunciation, as well as ecstatic abundance ("Everything in moderation, even excess").

We practice various forms of sitting and moving meditation to (try to) bring focus to our minds and give us insights that we can apply to our individual or group creative work. In turn, the process of art-making benefits our progress along our personally-chosen psycho-spiritual paths, which may differ greatly from monk to monk.

At an institutional level, we focus on practice as much as on product. We consider ourselves an alternative form of art institution, and thus prize new ways of thinking about art: as devotion, as offering, as gift, as sacred act, as ritual, as individual expression of the divine, as teacher, as priest, as sacred text, and as sacrifice.

Similarly, we are an alternative economy: we practice resourcefulness and community in order to liberate artists from the prevailing economics of art that tends to turn it into a commodity.

We apply monastic forms of governance to our community living: we experiment with monastic rules and vows. For a few years, for example, many of us took yearlong "Artmonk Vows" to practice things like gratitude, presence, and resourcefulness.

We think about art as a lineage, or as a number of branching lineages, some of which are directly applicable to our vision of personal awakening and cultural transformation. Like other monastic traditions, we experiment with functional (rather than absolute) hierarchies. Each Art Monastery could have an Abbot or Abbess, as well as a spiritual director and an artistic director.

Like many monasteries, we are interested in being both removed from the world and actively engaged to make it better through service and hospitality. We have a commitment to the places we live. Like many traditions, we work with philosophical dialogue and debate to bring about conceptual artistic and philosophical/spiritual understanding.

This process is just a few years old, and we basically everything yet to learn. For example, even though we have occupied a few medieval monasteries in Italy, most of us have more experience with meditation retreats in Zen, Tibetan and Theravada Buddhist lineages than with the realities of western monastic life. And while the ideas for much of what we do are borrowed from other monastic sources, they often have to be recompiled from the ground up to be useful and relevant to us. How can we honor tradition, even as we borrow from and adapt it to our unique mission?

These are the kinds of creative problems artmonks love most.

The Utility of Human Atoms for the Paperclip Maximizer

TL;DR: use of humans' atoms will increase AI's expected utility by 0.00000000000000002 of total U.

The iconic example of the existential risk of superintelligence is the so-called paperclip maximizer, that is, a system which maximizes some random goal not aligned with human values. While a real paperclip maximizer is unlikely, convergent dynamic subgoals suggested by Omohundro, imply that almost any non-aligned AI will try to acquire as much computational power as possible, as well as other material resource. Thus, it will need the atoms of humans.

In this article, I will try to estimate the marginal utility of human atoms U¢ compared with total utility of all atoms' U for the AI.

I assume that:

A) Superintelligent AI which is interested in human atoms, is limited to the Solar System. The reasons for this is that it may need human atoms to build spaceships to travel to the nearest stars, and if it reaches the stars, the value of most human atoms will be lost.

B) Superintelligent AI has powerful nanotechnology, which could be used to do whatever it wants with atoms, like build large space ships, deconstruct planets down to their final atoms etc.

C) The only resource that AI is interested in is atoms—not energy, not information etc.

D) The utility of atoms' U is increasing linearly with increases in the available mass.

The weight of Solar System planets is 0.135% of the Sun's mass, which is 1.98855×10E30 kg, and the mass of all planets is 2.6E1027kg, where most of this mass is Jupiter. Earth's mass is 5.97237×10E24 kg.

The total mass of human beings on Earth is 70kg x 7.5x10E9 people = 5.25x10E11kg.

In that case, human atoms constitute marginal utility U'= 2x10E-16U. That is, use of the humans' atoms will increase the AI's utility by 0.00000000000000002 of total U.

But this is obviously flawed, as humans can't survive in a vacuum, they need the whole Earth!

Imagine that AI wants to deconstruct the Earth for its atoms, but also decided to preserve human lives. It has two options:

1) Upload everybody into smaller computers. Based on various estimations of the Launder limit, and the computational capacity of human brains, the size of these computers will be different, but they could easily be 1000 times less than human bodies.

2) AI decides not to upload humans, but to build a space station, where humans can live approximately the same life as they do currently. As the typical weight of a human house is something like 10 tons, and assuming very effective nanotech, such a space station may require hardware weighing 1000 times more kg for every human kg (or perhaps even less). It will weigh 5.25x10E14kg.

Option (1) is a million times more economical than the option (2) for the AI. But even in the case of option (2), the marginal utility of human atoms U'= 2x10E-13U.

If the number of the atoms linearly translates into the speed of the start of the galactic colonization (Armstrong & Sandberg, 2013), and AI needs 1 billion seconds (30 years) to convert all the Solar System planets into space ships, the delay caused by preserving humans on a space station will be around 0.2 milliseconds.

Given Bostrom's astronomical waste idea (Bostrom, 2003), that may be not small after all, as it will increase the sphere of the AI's reach by 150 km, and after billions of years it will correspond to very large volume (assuming the size of the universe is like 10E21 light milliseconds, and the number of stars in it is around 10E24, times an economy of 0.2 milliseconds, could mean gain of more than 1000 stars, equal to hundreds of solar masses.)

Even Friendly AI may deconstruct humans for their atoms in the AI's early stages, and as such sacrificy will translate in the higher total number of sentient beings in the universe at the end.

In the another work (Turchin, 2017), I suggested that "price" of human atoms for AI is so infinitely small, that it will not kill humans for their atoms, if it has any infinitely small argument to preserve humans. Here I suggested more detailed calculation.

This argument may fail if we add the changing utility of human atoms over time. For early AI, Earth's surface is the most available source of atoms, and organic matter is the best source of carbon and energy (Freitas, 2000). Such an AI may bootstrap its nanotech infrastructure more quickly if it does not care about humans. However, AI could start uploading humans, or at least freezing their brains in some form of temporary cryostasis, even in the very early stages of its development. It that case AI may be acquire most human atoms without killing them.

Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, *89*, 1–13.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308–314.

Freitas, R. (2000). *Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations*. Foresight Institute Technical Report.

Turchin, A. (2017). Messaging future AI. Retrieved from https://goo.gl/YArqki

Active vs Passive Distraction

When life gets difficult, it can often become tempting to distract yourself from your own thoughts and feelings. Loved ones may scold us for this, telling us that we should confront our feelings, because distracting ourselves doesn't accomplish anything. However, distraction is an important part of distress tolerance. Imagine Jill, a young woman who, in typical romantic comedy fashion, gets fired and dumped on the same day. Sure, distracting herself isn't going to help her find a new job, but binge-watching her favorite TV series would give Jill time to let her emotions around the problem cool down before she decides to start working on it.

Of course, different kinds of people may require different kinds of distraction. For Jill, a more passive distraction like watching TV is perfect; she may not have the energy to do much else. However, her now-ex Bob is worried about how awkward their break-up is going to make things with their mutual friend group. If he sat down to watch TV, he would still be worrying about it. Bob needs a distraction that he can throw himself into, something that takes so much effort and focus that he can't spend any time worrying. So instead, he decides to pick up his guitar and practice a new song he's been trying to learn. Bob opts for a more active distraction.

The examples above not only show the difference between passive and active distraction, but also how to match what kind of distraction you need to your mood. For a lot of people, depression makes it hard to do things, so a more passive distraction may be preferred. Jill would not have had the energy to practice an instrument while depressed. On the other side, anxiety often manifests as a sort of nervous energy, which can easily be redirected. Of course, this isn't the case for everyone. Some people find anxiety paralyzing, so they may prefer a passive distraction despite being anxious. The important thing here is that the activeness and passivity of your distraction match your energy level.

What some people find to work as a distraction, others might find works better as emotional regulation. Maybe the real reason Jill didn't want to play music is that it wouldn't distract her at all, and would instead remind her of Bob even more! A few days after the break-up, Jill may still be feeling sad and missing him, so she sits down at her piano and plays the saddest song she knows. Here she is not using music to distract her from her feelings. On the contrary, she is leaning into her emotions, and letting them out. The song she plays could be the same exact song that Bob was trying to learn on his guitar; the important thing here is the approach. Jill is choosing an activity that matches her mood. If she was angry about the breakup, she might play an angry song, or go to a kickboxing class. If she was sad, but didn't want to play music, she might write a long letter to Bob, then shred it.

Not only is distraction anecdotally helpful, but research has been done to track symptoms in those who do and don't use the technique. A Swedish study showed that patients admitted to the hospital after a car accident were less likely to develop PTSD if they played Tetris within a few hours of admission. Some of the core symptoms of PTSD involve recurring thoughts, intrusive memories, and flashbacks, and it is thought that distracting a person from ruminating over the event blocks this pattern from forming. Those who distracted in this study had fewer intrusive memories in the week following, and these intrusive memories diminished faster. An earlier Oxford study showed that in a case of simulated trauma, playing Tetris was the best of three options, with taking an online trivia quiz as worse than doing nothing. This indicates

that choosing the wrong distraction technique (in this case, a passive one rather than an active one) can actually be hurtful. In this case, it makes sense that a failed distraction could actually train the brain to ruminate even while occupied. A distraction must be sufficiently distracting without being overwhelming.

This is not to say that distraction is always the best technique to use in a given situation. Distraction works best as distress tolerance, and is not a replacement for emotional regulation. As discussed above, the two work differently. Distracting yourself instead of using emotional regulation can lead to unprocessed or buried thoughts and feelings. In the moment, it can be hard to tell whether distracting yourself is helping or not. Different people have different tells (losing track of time, forgetting to do something important like eat lunch, etc) but generally, distraction should make you feel better, not worse. If you feel worse after a period of distraction, it may not be the right technique to be using. So next time you are feeling overwhelmed and need to veg out, go ahead! Just take the time to think about the kind of distraction you need first.

Knowledge is Freedom

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

[Epistemic Status: Type Error]

In this post, I try to build up an ontology around the following definition of knowledge:

To know something is to have the set of policies available to you closed under conditionals dependent on that thing.

You are an agent G, and you are interacting with an environment e in the set E of all possible environments. For each environment e, you select an action a from the set A of available actions. You thus implement a policy $p \in A^E$. Let $P \subseteq A^E$ denote the set of policies that you could implement. (Note that A^E is the space of functions from E to A.)

If you are confused about the word "could," that is okay; so am I.

A fact (F, ϕ) about the enviornment can be viewed as a function $\phi : E \to F$ that partitions the set of environments according to that fact. For example, for the fact "the sky is blue," we can think of F as the set $\{\top, \bot\}$ and ϕ as the function that sends worlds with a blue sky to the element $\top$ and sends worlds without a blue sky to the element $\bot$. One example of a fact is (E, id) which is the full specification of the environment.

A conditional policy can be formed out of other policies. To form a conditional on a fact (F, ϕ) we start with a policy for each element of F. We will let c(f) denote the policy associated with $f \in F$, so $c : F \to A^E$. Given this fact and this collection of policies, we define the conditional policy $p_c : E \to A$ given by $e \mapsto c(\phi(e))(e)$.

Conditional policies are like if statements in programming. Using the fact "the sky is blue" from above, we can let k_r be the policy that pushes a red button regardless of its environment and let k_g be a policy that pushes a green button regardless of its environment. If $c(\top) = k_r$ and $c(\bot) = k_g$, then p_c is the policy that pushes the red button if the sky is blue, and pushes big green button otherwise.

Now, we are ready to define knowledge. If P is the set of policies you could implement, then you know a fact (F, ϕ) if P is closed under conditional policies dependent on F. (i.e. Whenever $c : F \to P$, we have $p_c \in P$.) Basically, we are just saying that your policy is allowed to break into different cases for different ways that the fact could go.

Self Reference

Now, let's consider what happens when an agent tries to know things about itself. For this, we will consider a naturalized agent, that is part of the environment. There is a fact (A, action) of the environment that says what action the agent takes, where A is again the set of actions available to the agent, and action is a function from E to A that picks out what action the agent takes in that environment. Note that action is exactly the agent's policy, but we are thinking about it slightly differently.

So that things are not degenerate, let's assume that there are at least two possible actions a and b in A, and that P contains the constant policies k_a and k_b that ignore their evironment and always ouptut the same thing.

However, we can write down an explicit policy that the agent cannot implement: the policy where the agent takes action b in environments in which it takes action a, and takes action a in environments in which it does not take action a. The agent cannot implement this policy, since there are no consistant environments in which the agent is implementing this policy. (Again, I am confused by the coulds here, but I am assuming that the agent cannot take an inherently contradictory policy.)

This policy can be viewed as a conditional policy on the fact (A, action). You can construct it as p_c, where c is the function that maps a to k_b and everything else to k_a.

The fact that this conditional policy cannot be in P shows that the agent cannot by our definition know its own action.

Partial Knowledge

As seen above, there are limits to knowledge. This makes me want to aim lower and think about what types of partial knowledge can exist. Perhaps an agent can interact with a fact in nontrivial ways, while still not having complete knowledge defined above. Here, I will present various ways an agent can have partial knowledge of a fact.

In all of the below examples we will use a fact $(\{1,2,3,4\},\phi)$ about the environment that can take on 4 states, an action that can take on four values $A = \{ac, ad, bc, bd\}$, and we assume that the agent has access to the constant functions. Think about how all of these types of partial knowledge can be interpreted as changing the subet $P \subseteq A^E$ in some way.

Knowing a Coarser Fact: The agent could know a fact that has less detail than the original fact, for example the agent could know the parity of the fact above. This would mean that the agent can choose a policy to implement on worlds sent to 1 or 3, and another policy to implement on worlds sent to 2 or 4, but cannot necessarily use any more resolution.

Knowing a Logically Dependent Fact: The agent could, for example, know another fact $(\{1,2,3,4,\bot\},\phi')$ with the property that $\phi'(e) = \phi(e)$ whenever $\phi'(e) \neq \bot$. The agent can safely do policies when it knows it is in states 1 through 4, but it also might be in a state of uncertainty, and know the environment is $\bot$.

Knowing a Probabilistically Dependent Fact: The agent could, for example, know another fact $(\{1,2,3,4\},\phi')$, which is almost the same as the original fact, but is wrong in some small number of environments. The agent cannot reliably implement functions dependent on the original fact, but can correlate its action with the original fact by using this proxy.

Learning a Fact Later in Time: Imagine the agent has to make two independent actions at two different times, and the agent learns the fact after the first action, but before the second. In the above example, the first letter of the action, a or b, is the first action, and the second letter, c or d, is the second action. The policies are closed under conditionals as long as the different policies in the conditional agree on the first action. This is particularly interesting because it shows how to think of an agent moving through time as a single timeless agent with partial knowledge of the things that it will learn.

Paying Actions to Learn a Fact: Similar to the above example, imagine that an agent will learn the fact, but only if it chooses a in the first round. This corresponds to being closed under conditionals as long as all of the policies always choose a in the first round.

Paying Internal Resources to Learn a Fact: Break the fact up into two parts: the parity of the number, and whether the numer is greater than 2. Imagine an agent that is in an epistemic state such that it could think for a while and learn either of these bits, but cannot learn both in time for when it has to take an action. The agent can depend its policy on the parity or the size but not both. Interestingly, this agent has

strictly more options than an agent that only knows the parity, but technically does not fully know the parity. This is because adding more options can take away the closure property on the set of policies.

Other Subsets of the Function Space: One could imagine for example starting with an agent that knows the fact, but specifying one specific policy that the agent is not allowed to use. It is hard to imagine this as an epistemic state of the agent, but things like this might be necessary to talk about self reference.

Continuous/Computable Functions: This does not fit with the above example, but we could also restrict the space of policies to e.g. computable or continuous function of the environment, which can be viewed as a type of partial knowledge.

Confusing Parts

I don't know what the coulds are. It is annoying that our definition of knowledge is tied up with something as confusing as free will. I have a suspicion, however, that this is necessary. I suspect that our trouble with understanding naturalized world models might be coming from trying to understand them on their own, when really they have a complicated relationship with decision theory.

I do not yet have any kind of a picture that unifies this with the other epistemic primitives, like probability and proof, and I expect that this would be a useful thing to try to get.

It is interesting that one way of thinking about what the coulds are is related to the agent being uncertain. In this model, the fact that the agent could take different actions is connected to the agent not knowing what action it takes, which interestingly matches up with the fact in this model, if an agent could take multiple actions, it can't know which one it takes.

It seems like an agent could effectively lose knowledge by making precommitments not to follow certain policies. Normal kinds of precommitments like "if you do X, I will do Y" do not cause the agent to lose knowledge, but the fact it can in theory is weird. Also, it is weird that an agent that can only take one action vacuously knows all things.

It seems like to talk about knowing what you know, you run into some size problems. If the thing you know is treated as a variable that can take different values, that variable lives in the space of subsets of functions from environments to actions, 2^{A^E} which is much larger than E. I think to talk about this you have to start out restricting to some subset of functions from the beginning, or some subset of possible knowledge states.

How I see knowledge aggregation

After reading the Arbital postmortem, I remembered some old ideas regarding a tool for claim and prediction aggregation.

First, the tool would have the basic features. There would be a list of claims. Each claim is a clear and concise statements that could be true or false, perhaps with a short explanation. For each claim, the users could vote on its likelihood. All these votes would be aggregated into a single number for each claim.

Second, the tool would allow the creation of composite claims by combining two existing claims. In particular, a conditional claim *IF B THEN A* would represent the conditional probability $P(A|B)$. For every claim, it should be easy to find the conditionals it participates in, or the claims it is composed of. Conditionals are voted on same as simple claims (I would even consider a version where *only* conditionals are voted on).

Third, the tool would understand the basic probability laws and use this to direct the users' attention. For example, if three claims don't satisfy the law $P(A|B)\ P(B) < P(A)$, users might be alerted about this error. On the other hand, if $P(A|B) = P(B|A) = 1$, the two claims might be merged or one could be discarded, to reduce the clutter.

Fourth, given a claim the tool might collect every other claim that supports it, follow every chain of argument and assemble them into a single graph, or even a semi-readable text, with the strongest arguments and counterarguments most visible.

Let's consider a possible workflow. Suppose you browse the list of claims, and find a ridiculous claim X assigned a high likelihood. You could just vote to decrease the likelihood and perhaps leave an offensive comment, however this is unlikely to have much effect. Instead you could find a convincing counterargument Y, then add both $P(Y) = 1$ and $P(X|Y) = 0$ to the list of claims. Now other users would be notified of the resulting inconsistency and would reply by voting on one of these claims, changing their vote on X, or by creating additional arguments that contradict Y or support X. In turn you would attack these new arguments, eventually creating a large graph of reasoning. Perhaps at some point there would be enough general claims that people from different debates could reuse some, instead of duplicating them, and only create new conditional claims, making the graph dense.

I took some time to write a small prototype. It is initialized with some AI related claims, and it implements first, second, and a tiny bit of third paragraph. Now for some meta. The prototype is a single page app, using angular1, backed by nodejs and mongodb. The actual features took between 1 and 2 hours to write. The backend and deployment took a couple more hours, largely because I hadn't done that in a while. Therefore I think that it's quite feasible to make similar prototypes for other ideas. Is there any value in it though?

Bug Hunt 2

This is part 11 of 30 of Hammertime. Click here for the intro.

CFAR has an underlying mantra “adjust your seat”: systematically modify every technique and class to fit your personal situation. It’s common sense nowadays that different things work for different people, but the extent to which is true still constantly surprises me. (Kierkegaard had a fun take on adjusting your seat which he called the Method of Rotation.)

If you wish to partake in Hammertime, feel free to adjust your seat as much as necessary. Draw out the practice of instrumental rationality over a longer period of time, pick and choose the methods that appeal to you, and scale them to your time constraints.

Hammertime: The Second Cycle

Hammertime is about cultivating a tiny number of powerful techniques for solving a huge variety of problems. In the second cycle, we will revisit and upgrade the tools we introduced in the first, and apply them to tougher problems:

1. Bug Hunt
2. Yoda Timers
3. TAPs and Reinforcement Learning
4. Design
5. CoZE
6. Focusing
7. Cruxes
8. Goal Factoring
9. Internal Double Crux
10. Self-Trust

The new ideas we will be introducing in the second half are devoted to developing higher levels of introspection and self-honesty, to figure out your true motivations and aversions, and what to do about them.

Before each post in the second cycle, take a moment to review its predecessor.

Day 11: Bug Hunt 2

Previously: Day 1.

Noticing your bugs continues to be our single most powerful technique. Training noticing involves lateral thinking, attention to detail, and self-honesty. Today, we focus on three high-level ways in which human beings systematically err.

Setup

First, review your Bug List from Day 1 and update it.

For each of the next three mini-essays: read it over, then set a Yoda Timer for five minutes and brainstorm as many bugs as you can during that time.

1. Identity

Paul Graham wrote Keep Your Identity Small. Being attached to your identity can often constrain your growth.

Rather than making an impartial decision on what kind of person to be, people often extrapolate their identity (and morality) from their previous actions. A friend of mine calls this coprolite: fossilized and over-fitted beliefs that originate from early childhood. Are you neat or messy, stingy or generous, introvert or extrovert, conscientious or agreeable, idealistic or cynical, engineer or artist, vim or emacs? Do you look down on people for being the other way? Take moment to notice all the traits you're attached to, think about why you're attached to them, and consider the benefits of their opposites.

Personalities are many-faceted, and you may not even understand your true motives, fears, or skills. Do your stated preferences agree with your revealed preferences? Do your aliefs differ from your beliefs? Do people systematically judge you to be different from your self-image? Do you often surprise yourself in terms of what you enjoy, excel at, or are anxious about?

It's useful to think of personality growth as expansion rather than change. An introvert grows by learning how to navigate social scenes. An extrovert grows by reclaiming her capacity to be alone. Instead of asking what you would change about yourself, ask what you would add to the toolbox.

2. Pica

Pica is an eating disorder in which people crave food that don't fulfill the need behind that craving; the folklore example is gnawing on ice to satisfy a mineral deficiency. Experiential pica is any craving which doesn't fulfill the need behind it.

My top three addictions in high school were all experiential pica.

The first addiction was romantic novels and shows of a tragic nature, which served as vulnerability and sacrifice porn. I had intricate daydreams in multiple languages of love and loss.

The second addiction was RPG games, which served as improvement porn. In Diablo III, the Gem of Ease that boosts your leveling speed on all future characters to go from 1 to 70 in about an hour; I'd start a new character every couple months to get watch the level up messages roll in. MOBAs are perhaps the worst offender in this regard, taking your character from level 1 to fully equipped level 18 every single game.

The third addiction was just …

I know these are pica because the first and third cravings largely subsided when I entered a committed relationship, and the second when I started seriously working on self-improvement.

Lent is a good time to look for your pica. Are there any habits, cravings, or addictions you don't understand and/or try hard to cut? If they're pica, you're applying effort at the wrong angle. Figure out the unmet need, meet it, and the pica will automatically subside.

3. Ambition

I've been jogging casually for about fifteen years. Until last year, it's been uniformly awful. You'd think you'd get used to running four miles after doing it twice a week for a decade. You'd be wrong.

Then, I decided to aim at something.

I thought: *I'm going to train for a seven minute mile*.

My heart replied: *Oh, ok, that's kind of invigorating.*

Then I thought: *I'll train for a six minute mile.*

My heart: *Woo baby, let's do this!*

Then I thought: *A five minute mile.*

My heart: *HAHAHAHAHAHAHAHAHA…*

I ran for over a decade with next to no improvement. Last month I ran a seven minute mile after two months of training for an impossible goal. These days I look forward to running.

I've been blogging casually for about five years. Until last year, it's been a drag. You'd think you'd get better at writing by putting up two posts a month for a year or two. You'd be wrong.

Then, I decided to aim at something.

Me: *I'll try blogging once a week.*

My heart: *Oh, ok, that's nice.*

Me: *I'm going to blog every other day.*

My heart: *Now we're getting somewhere.*

Me: *I'm going to blog every day for a year, and write better than Eliezer Yudkowsky by the end of it.*

My heart: *HAHAHAHAHAHAHAHAHA…*

There's a level of ambition that pushes you to operate at maximal efficiency, that twists your heart with adrenaline just to think about. In every pursuit, aim at a target so high it feels immodest to whisper in an empty room.

List your goals now. Keep doubling them in difficulty until your heart bends over in hysterical laughter at the very thought.

Daily Challenge

State your greatest ambition: the one that feels most subjectively immodest.

The map has gears. They don't always turn.

Follow up to Toward a New Technical Explanation of Technical Explanation.

Confusion is in the map, not the territory. The same goes for predictability, surprise, mysteriousness, weirdness, and so on. Yet, sometimes, we say "no, actually, that's quite surprising" after initially not being surprised -- and it has a meaning. Or, sometimes we say "ah, right, it *is* obvious" after being confused and needing to think for a while -- and it has meaning.

Though they may be phrased as map/territory errors, statements like this can refer to the amount of surprise or confusion etc which a specific gears-level model has. A step in a proof may be "obvious" because it is an elementary application of a common method, even if it takes time for a fallible human to see it. A physicist may one day come to realize that a commonplace event is "surprising" after being not-surprised by the event dozens of times, because suddenly they've compared it to a model from physics, and found the event difficult to account for.

Our brain has a big messy model of reality, but it is often trying to simulate smaller, cleaner, more justified, more objective, and hopefully better models. When we do things like this, we are speaking from those models.

Some conceptual highlights from “Disjunctive Scenarios of Catastrophic AI Risk”

My forthcoming paper, “Disjunctive Scenarios of Catastrophic AI Risk”, attempts to introduce a number of considerations to the analysis of potential risks from Artificial General Intelligence (AGI). As the paper is long and occasionally makes for somewhat dry reading, I thought that I would briefly highlight a few of the key points raised in the paper.

The main idea here is that most of the discussion about risks of AGI has been framed in terms of a scenario that goes something along the lines of “a research group develops AGI, that AGI develops to become superintelligent, escapes from its creators, and takes over the world”. While that is one scenario that could happen, focusing too much on any single scenario makes us more likely to miss out alternative scenarios. It also makes the scenarios susceptible to criticism from people who (correctly!) point out that we are postulating very specific scenarios that have lots of burdensome details.

To address that, I discuss here a number of considerations that suggest *disjunctive* paths to catastrophic outcomes: paths that are of the form “A *or* B *or* C could happen, and any one of them happening could have bad consequences”.

Superintelligence versus Crucial Capabilities

Bostrom’s *Superintelligence*, as well as a number of other sources, basically make the following argument:

1. An AGI could become superintelligent
2. Superintelligence would enable the AGI to take over the world

This is an important argument to make and analyze, since superintelligence basically represents an extreme case: if an individual AGI may become as powerful as it gets, how do we prepare for that eventuality? As long as there is a plausible chance for such an extreme case to be realized, it must be taken into account.

However, it is probably a mistake to focus only on the case of superintelligence. Basically, the reason why we are interested in a superintelligence is that, by assumption, it has the cognitive capabilities necessary for a world takeover. But what about an AGI which also had the cognitive capabilities necessary for taking over the world, and only those?

Such an AGI might not count as a superintelligence in the traditional sense, since it would not be superhumanly capable in *every* domain. Yet, it would still be one that we should be concerned about. If we focus too much on just the superintelligence case, we might miss the emergence of a “dumb” AGI which nevertheless had the *crucial capabilities* necessary for a world takeover.

That raises the question of what might be such crucial capabilities. I don’t have a comprehensive answer; in my paper, I focus mostly on the kinds of capabilities that

could be used to inflict major damage: social manipulation, cyberwarfare, biological warfare. Others no doubt exist.

A possibly useful framing for future investigations might be, “what level of capability would an AGI need to achieve in a crucial capability in order to be dangerous”, where the definition of “dangerous” is free to vary based on how serious of a risk we are concerned about. One complication here is that this is a highly contextual question – with a superintelligence we can assume that the AGI may get basically omnipotent, but such a simplifying assumption won’t help us here. For example, the level of offensive biowarfare capability that would pose a major risk, depends on the level of the world’s defensive biowarfare capabilities. Also, we know that it’s possible to inflict enormous damage to humanity even with just human-level intelligence: whoever is authorized to control the arsenal of a nuclear power could trigger World War III, no superhuman smarts needed.

Crucial capabilities are a disjunctive consideration because they show that superintelligence isn’t the only level of capability that would pose a major risk: and there many different combinations of various capabilities – including ones that we don’t even know about yet – that could pose the same level of danger as superintelligence.

Incidentally, this shows one reason why the common criticism of “superintelligence isn’t something that we need to worry about because intelligence isn’t unidimensional” is misfounded – the AGI doesn’t need to be superintelligent in *every* dimension of intelligence, just the ones we care about.

How would the AGI get free and powerful?

In the prototypical AGI risk scenario, we are assuming that the developers of the AGI want to keep it strictly under control, whereas the AGI itself has a motive to break free. This has led to various discussions about the feasibility of “oracle AI” or “AI confinement” – ways to restrict the AGI’s ability to act freely in the world, while still making use of it. This also means that the AGI might have a hard time acquiring the resources that it needs for a world takeover, since it either has to do so while it is under constant supervision by its creators, or while on the run from them.

However, there are also alternative scenarios where the AGI’s creators voluntarily let it free – or even place it in control of e.g. a major corporation, free to use that corporation’s resources as it desires! My chapter discusses several ways by which this could happen: i) economic benefit or competitive pressure, ii) criminal or terrorist reasons, iii) ethical or philosophical reasons, iv) confidence in the AI’s safety, as well as v) desperate circumstances such as being otherwise close to death. See the chapter for more details on each of these. Furthermore, the AGI could remain theoretically confined but be practically in control anyway – such as in a situation where it was officially only giving a corporation advice, but its advice had never been wrong before and nobody wanted to risk their jobs by going against the advice.

Would the Treacherous Turn involve a Decisive Strategic Advantage?

Looking at crucial capabilities in a more fine-grained manner also raises the question of *when* an AGI would start acting against humanity’s interests. In the typical superintelligence scenario, we assume that it will do so once it is in a position to achieve what Bostrom calls a Decisive Strategic Advantage (DSA): “a level of

technological and other advantages sufficient to enable [an AI] to achieve complete world domination". After all, if you are capable of achieving superintelligence and a DSA, why act any earlier than that?

Even when dealing with superintelligences, however, the case isn't quite as clear-cut. Suppose that there are two AGI systems, each potentially capable of achieving a DSA if they prepare for long enough. But the longer that they prepare, the more likely it becomes that the other AGI sets its plans in motion first, and achieves an advantage over the other. Thus, if several AGI projects exist, each AGI is incentivized to take action at such a point which maximizes its overall probability of success – even if the AGI only had rather slim chances of succeeding in the takeover, if it thought that waiting for longer would make its chances even worse.

Indeed, an AGI which defects on its creators *may not be going for a world takeover in the first place*: it might, for instance, simply be trying to maneuver itself into a position where it can act more autonomously and defeat takeover attempts by other, more powerful AGIs. The threshold for the first treacherous turn could vary quite a bit, depending on the goals and assets of the different AGIs; various considerations are discussed in the paper.

A large reason for analyzing these kinds of scenarios is that, besides caring about existential risks, we also care about *catastrophic* risks – such as an AGI acting too early and launching a plan which resulted in "merely" hundreds of millions of deaths. My paper introduces the term Major Strategic Advantage, defined as "a level of technological and other advantages sufficient to pose a catastrophic risk to human society". A catastrophic risk is one that might inflict serious damage to human well-being on a global scale and cause ten million or more fatalities.

"Mere" catastrophic risks could also turn into existential ones, if they contribute to *global turbulence* (Bostrom et al. 2017), a situation in which existing institutions are challenged, and coordination and long-term planning become more difficult. Global turbulence could then contribute to another out-of-control AI project failing even more catastrophically and causing even more damage

Summary table and example scenarios

The table below summarizes the various alternatives explored in the paper.

AI's level of strategic advantage

- Decisive
- Major

AI's capability threshold for non-cooperation

- Very low to very high, depending on various factors

Sources of AI capability

- Individual takeoff
    - Hardware overhang
    - Speed explosion
    - Intelligence explosion

- Collective takeoff
- Crucial capabilities
    - Biowarfare
    - Cyberwarfare
    - Social manipulation
    - Something else

- Gradual shift in power

Ways for the AI to achieve autonomy

- Escape
    - Social manipulation
    - Technical weakness

- Voluntarily released
    - Economic or competitive reasons
    - Criminal or terrorist reasons
    - Ethical or philosophical reasons
    - Desperation
    - Confidence
        - in lack of capability
        - in values

- Confined but effectively in control

Number of AIs

- Single
- Multiple

And here are some example scenarios formed by different combinations of them:

<u>The classic takeover</u>

(Decisive strategic advantage, high capability threshold, intelligence explosion, escaped AI, single AI)

The “classic” AI takeover scenario: an AI is developed, which eventually becomes better at AI design than its programmers. The AI uses this ability to undergo an intelligence explosion, and eventually escapes to the Internet from its confinement. After acquiring sufficient influence and resources in secret, it carries out a strike against humanity, eliminating humanity as a dominant player on Earth so that it can proceed with its own plans unhindered.

<u>The gradual takeover</u>

(Major strategic advantage, high capability threshold, gradual shift in power, released for economic reasons, multiple AIs)

Many corporations, governments, and individuals voluntarily turn over functions to AIs, until we are dependent on AI systems. These are initially narrow-AI systems, but continued upgrades push some of them to the level of having general intelligence. Gradually, they start making all the decisions. We know that letting them run things is risky, but now a lot of stuff is built around them, it brings a profit and they're really good at giving us nice stuff—for the while being.

The wars of the desperate AIs

(Major strategic advantage, low capability threshold, crucial capabilities, escaped AIs, multiple AIs)

Many different actors develop AI systems. Most of these prototypes are unaligned with human values and not yet enormously capable, but many of these AIs reason that some other prototype might be more capable. As a result, they attempt to defect on humanity despite knowing their chances of success to be low, reasoning that they would have an even lower chance of achieving their goals if they did not defect. Society is hit by various out-of-control systems with crucial capabilities that manage to do catastrophic damage before being contained.

Is humanity feeling lucky?

(Decisive strategic advantage, high capability threshold, crucial capabilities, confined but effectively in control, single AI)

Google begins to make decisions about product launches and strategies as guided by their strategic advisor AI. This allows them to become even more powerful and influential than they already are. Nudged by the strategy AI, they start taking increasingly questionable actions that increase their power; they are too powerful for society to put a stop to them. Hard-to-understand code written by the strategy AI detects and subtly sabotages other people's AI projects, until Google establishes itself as the dominant world power.

This blog post was written as part of work for the Foundational Research Institute.

Whose reasoning can you rely on when your own is faulty?

None of us are perfect reasoners. None of us have unlimited information. Sometimes other people are more correct than we are. This is an obvious thing we all *know* but may not *practice* .

Below are some concrete questions you can think about that come at this problem from a very specific path of trust (There are a lot of other ways to come at this). Note, that when I say "trust" I AM NOT NOT NOT saying that you should allow other people's viewpoints to completely supercede your own. The trust I am implying is more along the lines of something like "this would cause me to pause and consider."

Who do I trust to be a solid representation of an alternative viewpoint?

The most frequent arguments we hear for viewpoints different from are own are often not the *strongest* arguments. Who do you trust to present well-reasoned, not mindkilled arguments for a differing viewpoint, while not trying to persuade you with ideological rhetoric.

If I am in the throes of emotion, who could tell me that I wasn't behaving "rationally", AND (ideally) that information would actively cause me to change my behavior?

Particularly if you know you often behave in ways you don't endorse when you are angry, or starry-eyed, or perhaps under the influence of illicit substances... Is there someone who could say "You need to stop what you're doing right now." For me, there are occasional times that I am angry-and-irrational enough that I explicitly seek out an outside view as a check on my reasoning and actions. Any Friend I Trust is going to be a better judge of my actions at that time than I myself would be. Later, when I'm calmer, I can revisit and make sure I still endorse those conclusions.

Who would I most trust with end-of-life care?

This is a situation it is good to prepare for in advance because you may be completely unable to effect the decision by the time it comes up. So in this case you actually ARE putting important personal decisions completely in someone else's hands. My person is actually a former partner that I haven't spoken to in over a year. But we've had explicit discussions about our preferences (which are very similar), and I trust him to not be swayed by things like "What will other people think?" or "I need to prove my love by not letting go."

Next, do this in reverse. Think of some specific friends, and for each friend think about what they, specifically, as individuals, have to offer.

What is the sort of thing Alice In Particular is likely to get more correct than you? How much are you willing to pause, reflect, and potentially act on their input? Below are some axes of trust you can think about with each individual?

(It could also be interesting to do this with some people you don't particularly like or trust. Is there something you would still trust them with? Maybe they have some

domain knowledge and if they said "Your math is wrong" you would update enough to go back and check your work or ask for clarification.)

Axes of trust:

Do they have domain expertise that you don't have? Are they generally more intelligent or rational? Do they have a different but useful worldview? Do they have past experience you don't have access to (e.g. coming from another culture)? Do their incentives and goals align with yours? Do they care a lot about you? Do they have an accurate model of you and your options/ limitations?

If anyone has more axes or questions of trust to consider, I would love to hear them.

Hammertime Intermission and Open Thread

This post marks the end of the first cycle of Hammertime. Click here for intro.

Hammertime will return on Monday 2/19.

I want to close off the first cycle with some thoughts, and designate a place for discussion about the future of this sequence.

Discussion Topics

1. Sequences: Yea or Nay?

I've always felt that sequences are a valuable way to organize deeper thoughts and drive home a few central messages from several perspectives. However, the current format and culture on LW seem to radically favor short, independent chunks. (There is also the obvious problem that Sequence construction is not working.)

I've been posting daily for a while now but when I shifted from individual posts to a sequence, average Karma immediately dropped by a factor of about 2. It's possible that people don't bother upvoting the same sequence, or that my writing quality dropped, but if this is real signal that many more people would read a sequence if they are marketed as individual thoughts (and WordPress stats suggest this as well), that might be reason for me to stop writing sequences in the future, or at least collect sequences together only after they're complete.

Possible actionable for meta: make posts in a Sequence share karma and/or a single slot on frontpage.

2. Repeat or Explore?

My original intention was to review 10 topics over three cycles, building up in the difficulty of problems solved. I think I will definitely return to and expand on several of the techniques we've seen already, but also add more topics. If people have favorite techniques (and hopefully references) they'd like to see in Hammertime, post them here.

3. Monotonicity of Progress

A big goal of mine is to solve the "Rationalist Uncanny Valley," where beginning rationalists get worse at life before they get better. I can't believe that this has to be the case; it seems to be symptomatic of a larger failure to develop the proper curriculum. I would like progress on rationality to be monotone – is there a good reason this should be difficult? It'd be great if we could compile a central list of "uncanny valley" failure modes.

Confidence Confusion

> “Captain, if you steer the Enterprise directly into that black hole, our probability of surviving is only 2.234%” Yet nine times out of ten the Enterprise is not destroyed. What kind of tragic fool gives four significant digits for a figure that is off by two orders of magnitude?

~Why truth? And…

This post poses a basic question about probabilities that I’m confused about. Insight would be appreciated.

It’s inspired by a quick skim I gave to Proofiness, which argues that precise numbers are a powerful persuasion technique.

The Dinosaur Market

One of the CFAR prediction markets was: *A randomly selected participant will correctly name seven real species of dinosaurs.*

I made a series of suspect Fermi estimate as follows:

Extrapolating from the current bids, half the participants will not bid in this market. If selected, they will just try to name as many dinosaurs as they can. 20% of participants will be able to.

Half the participants do bid in this market. Among those that bid high on the market, 70% will take the time to study dinosaurs and memorize seven. The other half who bid low will intentionally or unintentionally fail if they’re paying attention. I give them a 10% chance of success.

That comes out to a total of 30%.

There’s a 5% chance of anomalous situations such as one person caring enough to teach people or publicly post dinosaur names. In this case much higher chances of market evaluating to True, say 60%.

*I arrived at a probability estimate of .95 * .3 + .05 * .6 = .315.*

At this point, I felt *obligated* to round 31.5% to 30%, and so I bid 30 on the market instead of 31 or 32. Is there a valid reason to do so?

Precision is Confidence

I’ve been focusing on my aversion to report high-precision numbers, even if I believe them to be closer to the truth. When I report 31.5%, I feel more confident than I am.

Teasing out what it means for 31.5% to be more confident 30% the issue is that any number comes with an implicit confidence interval based on the number of significant figures. 30 really means , whereas 31.5 really means .

In the absence of explicitly reporting confidence intervals around every probability estimate, 30 thus feels like a more honest report of my actual beliefs. Despite the simultaneous fact that I would buy at any price less than 31 and sell at any price over 32.

While explicitly reporting confidence intervals solves the issue – I'd rather say instead of – this strategy seems impractical and carries its own signalling problems.

A Thought Experiment

A large part of my aversion to putting precise numbers on beliefs is a result of the type of error above: in a social setting you cannot just say what you mean, and in particular the number of significant figures also signifies a level of confidence/information. This seems to be the norm: almost every prediction Scott Alexander makes is a multiple of 5 or 10.

Here's a thought experiment:

Albert and Betty are astronauts sent to study a mysterious coin on Europa and independently transmit short messages back to Earth about the coin's bias. Albert finds the coin successfully and flips it a thousand times, seeing 531 heads and 469 tails. He concludes the coin is fair and reports 50%.

Betty's landing capsule collides with a giant teapot in upper orbit and lands several hundred miles away from target. She still has to report her beliefs about the mysterious coin. She has two choices:

1. *Report 50%, because that's her prior in the absence of information. After all, probability is in the mind. If she does this, however, Albert's higher confidence is lost in transmission.*
2. *Report nothing to convey the fact that she has no information. Any other rationalist can automatically compute her odds are 50-50 anyway – that's a shared prior.*

What would you do? What are the correct norms around sharing probabilities in conversation?

If social norms indeed dictate that significant figures transmit confidence, might it be deceptive to report 31.5 instead of 30 in conversation about the dinosaur market?

Don't Condition on no Catastrophes

I often hear people say things like "By what date do you assign 50% chance to reaching AGI, conditioned on no other form of civilizational collapse happening first?" The purpose of this post is to make this question make you cringe.

I think that most people mentally replace the conditional with something like "if other forms of civilizational collapse were magically not a thing, and did not have to enter into your model." Further, I think this is the more useful question to discuss, as it makes it easier to double crux, or download other people's models. However, not everyone does this, and it is not the question being asked.

To illustrate the difference, consider Alice, who, if they ignored other civilizational collapse, would think that AGI arrival date is uniform over the next 100 years. However, they also think that if not for AGI, extinction level nuclear war will happen in the next 100 years, uniformly at random over the next 100 years. Alice is not concerned about any other catastrophes.

Alice has these two independent distributions on when each event will happen if counterfactually, the other were to magically not happen. However, the world is such that as soon as one of events happens, it causes the other event to not happen, because the world is made very different.

When asking about Alice's median AGI date, ignoring civilizational collapse, we would like to encourage her to say 50 years. However her median AGI date, conditional on no nuclear war happening first is actually 33 years. This is because conditioning on no nuclear war happening first biases towards AGI dates that are early enough to stop a counterfactual future nuclear war.

The form of the question I would like to ask Alice is as follows:

Take your distribution over ways the way the future can go, and sample a random future, f. If that future ends with nuclear war at time t, sample another world with the property that neither AGI nor any other catastrophe happens before time t. If that world ends with a non AGI catastrophe, redefine t to be the time of the catastrophe in that world, and repeat the process, until you get a world that ends with AGI, with no other catastrophe happening first. Use this as your new distribution over futures, and tell me the median AGI date.

Note that conditioning on no other catastrophe happening first is the same procedure, except when you sample a new future, you do not require that it has the property that neither AGI nor any other catastrophe happens before time t.

I don't have a good name for this alternative to conditioning, and would like suggestions in comments. You may notice a similarity between it and causal counterfactuals. You also notice a similarity between it and the thing you do to get Solomonoff Induction out of the Universal Semimeasure.

Pseudo-Rationality

Pseudo-rationality is the social performance of rationality, as opposed to actual rationality. Here are some examples:

- Being overly skeptical to demonstrate how skeptical you are
- Always fighting for the truth, even when you're burning more social capital than the argument is worth
- Optimising for charitability in discussions to the point where you are consistently being exploited
- Refusing to do any social signalling or ever bow to social norms to signal that you're above them
- Spending too much time reading rationality content or the kinds of things rationalists are interested in
- Adopting techniques like pomodoros or TAPs merely because all the cool (rationalist) kids are using them, instead of asking if they are really helping you
- Hating things like post-modernism because other rationalists hate them and not because you've actually thought about it for yourself (but yes, post-modernism is mostly incoherent)
- Over-analysing unimportant decisions so that you can prove you made the rational decision

Why does this happen? Status and social norms distort the way we see the world. Even if it doesn't fool everyone, it will fool some people. Or if it fools no-one, you'll at least fool yourself. Here are some thought patterns:

- All the other rationalists think I'm a good rationalist, surely I must be (all social incentive systems have loopholes)
- All the other rationalists do this, so it must be rational (can be applied even if you are doing it to a much higher degree)
- I am so much more rational than those other people who are wrong/bow to social norms/aren't at all skeptical (more rational does not equal rational)

Why did I write this post? Well, it seems the next thing you need after becoming a rationalist, is something to help you figure out if you're doing any of it wrong. I hope this helps, but let me know if I should add anything else to the list.

Reflection based on comments:

Where this gets complex is when you desire the successful social performance of rationality as a goal that holds up after reflection. Some people may value this to a level that seems excessive to most people and so may not be acting irrationally. More generally, it seems that every rationalist should value successfully performing rationality to some degree, even if only instrumentally. These considerations complicate discussions of what is or is not pseudo-rational, but do not invalidate the general concept as most often they are not in line with someone's considered values. Further, this concept has utility as identifying a pattern of behaviour that we might want to discourage as a community.

Footnotes:

This is very similar to Straw Vulcans except that Straw Vulcans are about how the media represents being logical/rational, while pseudo-rationality is broader and

includes misconceptions that may not be prevalent in the media. Another difference is that Straw Vulcans are about defending rationality/logic from being straw-manned, while pseudo-rationality is encouraging rationalitists to consider whether they are really as rational as they think they are.

Also see: Mythic values vs. folk values. Pseudo-rationality is very similar to folk values, pseudo-rationality is not about impressing other people, but about fooling yourself.

Status: Map and Territory

I'm here to add another angle to the discussion on social vs. objective truth (example). Here's an analogy for reasoning about status games and why people react so strongly against improper status moves:

Society is a collective consciousness. From Society's point of view, the status game is the map. Genuine competence (some combination of skill, virtue, and value) is the territory. The map is meant to track the territory.

Human instinctively play the status game; it's impossible to just say what you mean. The status game is built into people's verbal and nonverbal behaviors toward one another.

If the status game is a good map, you can decide who to befriend, admire, and chastise based simply on their status moves. You can figure out who best to ask for advice by the way they hold their arms. You can trust the beliefs of confident people without individually investigating each of their claims. The human brain opts into the status game by default to partake in all this free value.

If the accuracy of the status game is corrupted, the map loses all value. Trust breaks down and you have to rely on first principles.

There's an approved way of climbing the status ladder: acquiring genuine competence. Well-socialized individuals naturally play higher status as they become more competent in the relevant domain, since the connection between competence and status is built into their brains. Society approves: the map keeps fidelity to the territory.

There's an improper way of climbing the status ladder: playing status above your competence. Jordan Peterson's go-to example is serial killer Paul Bernardo in this prison interview. Note the minute-long interaction between Bernardo and the lawyer(?) on the right. Bernardo acts like a disappointed CEO lecturing a wayward and nervous underling.

Knowing the truth about the individuals involved, I have a visceral reaction against this status interaction: the map has detached from the territory. Even if Bernardo is speaking only literal truths, there's an instinct that screams he's *lying*.

I predict that the neural mechanisms for detecting truth from falsehood (i.e. whether the map corresponds to the territory) are closely related to the mechanisms for distinguishing proper and improper status moves (i.e. whether the status map corresponds to genuine competence). I predict that your negative reaction against lying feels similar to your negative reaction against improper status plays.

June 2012: 0/33 Turing Award winners predict computers beating humans at go within next 10 years.

In June 2012, the Association for Computing Machinery—a professional society of computer scientists, best known for hosting the prestigious ACM Turing Award, commonly referred to as the "Nobel Prize of Computer Science"—celebrated the 100th birthday of Alan Turing.

The event was attended by luminaries like, oh, in no particular order: Donald Knuth, Vint Cerf, Bob Kahn, Marvin Minsky, Judea Pearl, Ron Rivest, Adi Shamir, Leonard Adleman, and of extra relevance here; Ken Thompson, inventor of the UNIX operating system, co-inventor of the C programming language, and computer chess pioneer.

In all, some 33 Turing Award winners were scheduled to be in attendance. There were no parallel tracks or simultaneous panels going on.

Image credit: Joshua Lock

So today, randomly watching one of the panel debates on Youtube, I was amazed by the amusing / horrifying example of failure of foresight and predictive accuracy, by these world leading computer scientists, regarding the advancement of the state of the art in artificial intelligence.

In this case the as it pertains to the ancient board game "go".

https://www.youtube.com/watch?v=dsMKJKTOte0&t=54m

> "When does a computer crack go?"
>
> "And I will start by a 100 years, and then count down by ten year intervals."
>
> And [by] 90 years? I count about 4% of the audience. (...)

Perhaps my internet searching skills are weak, but as best as I can tell, the incident has not been noted other than a few bemused Youtube comments in the video linked above.

Given ten options, ten buckets in which to place their bet, world-leading experts in computer science, as a group, managed to perform much worse than one would expect given their vast and wide-ranging expertise.

Worse even, than one would expect of a group of complete ignorants.

Given ten options, one would expect one out of every ten to land on the right answer, if nobody knew anything and everyone made a blind guess.

Three years and three months later, three-time European champion Fan Hui, was defeated. Half a year after that, world champion Lee Sedol was defeated by AlphaGo.

Interestingly, Ken Thompson, to whom the query was first directed, starts his answer by sharing an experience, from a World Computer Chess Championship, around 1980. Participants were asked if and when computers would beat the world champion at the game of chess. And, he explains, everyone except him alone, had been exceedingly optimistic.

Thereby priming the audience in several ways for the informal poll which followed.

By demonstrating authority, by proving a track record of sorts, by providing an anchor of sorts, and by warning against optimism. (Thompson had predicted that computers would beat human champions at chess by 2011.)

> "Most of the predictions were like next year or five years.
> You know, way, way optimistic.
> If you look at Moore's law, on computers, and you look at the increase in speed. Ah, with strength, with speed, on computer chess,
> never is just, you know, you can't predict never.
> You just can't predict never.
> It had to happen."
>
> Moderator: "Do you think go then, is in the targets, of computers?"
>
> Ah no, I don't think go is in... Ahh... If I had to predict go, I'd predict way, way out.

And then, the audience was polled, by simple show of hands.

One member of the audience stood alone, red-faced, to laughter and ridicule from the most esteemed peers in his field.

One single member out of the whole audience got it right.

Links:

- MP4 Video and Panel description at ACM.org: https://dl.acm.org/citation.cfm?id=2322182
- Video by ACM on Youtube at: https://youtu.be/dsMKJKTOte0?t=54m
- ACM Turing Centenary celebration program:
- https://web.archive.org/web/20120617021752/http://turing100.acm.org:80/final*program/tcc* final_program.pdf
- ACM Turing 100 website: http://turing100.acm.org/ (dead link)
- https://www.flickr.com/photos/31039727@N02/7398725286/in/album-72157630186920410/
- AlphaGo on Wikipedia: https://en.wikipedia.org/wiki/AlphaGo