Best of LessWrong: January 2020

1. Reality-Revealing and Reality-Masking Puzzles
2. CFAR Participant Handbook now available to all
3. What cognitive biases feel like from the inside
4. Coordination as a Scarce Resource
5. 2018 Review: Voting Results!
6. Moral public goods
7. On hiding the source of knowledge
8. AI Alignment 2018-19 Review
9. Becoming Unusually Truth-Oriented
10. Hedonic asymmetries
11. Technology Changes Constraints
12. Of arguments and wagers
13. A rant against robots
14. Review: How to Read a Book (Mortimer Adler, Charles Van Doren)
15. Draining the swamp
16. Exploring safe exploration
17. Why Do You Keep Having This Problem?
18. The Road to Mazedom
19. Circling as Cousin to Rationality
20. Don't Double-Crux With Suicide Rock
21. Criticism as Entertainment
22. The Alignment-Competence Trade-Off, Part 1: Coalition Size and Signaling Costs
23. Studying Early Stage Science: Research Program Introduction
24. Homeostasis and “Root Causes” in Aging
25. What we Know vs. How we Know it?
26. Does GPT-2 Understand Anything?
27. Book review: Rethinking Consciousness
28. Material Goods as an Abundant Resource
29. Clarifying The Malignity of the Universal Prior: The Lexical Update
30. How Doomed are Large Organizations?
31. Appendix: how a subagent could get powerful
32. Use-cases for computations, other than running them?
33. Excitement vs childishness
34. How to Escape From Immoral Mazes
35. Subscripting Typographic Convention For Citations/Dates/Sources/Evidentials: A Proposal
36. The two-layer model of human values, and problems with synthesizing preferences
37. Disasters
38. [AN #83]: Sample-efficient deep learning with ReMixMatch
39. How to Identify an Immoral Maze
40. Moral uncertainty: What kind of 'should' is involved?
41. Healing vs. exercise analogies for emotional work
42. Using Vickrey auctions as a price discovery mechanism
43. Normalization of Deviance
44. Why a New Rationalization Sequence?
45. Update on Ought's experiments on factored evaluation of arguments
46. Safety regulators: A tool for mitigating technological risk
47. [AN #81]: Universality as a potential solution to conceptual difficulties in intent alignment
48. UML VII: Meta-Learning
49. Algorithms vs Compute
50. What is Life in an Immoral Maze?

Best of LessWrong: January 2020

1. Reality-Revealing and Reality-Masking Puzzles
2. CFAR Participant Handbook now available to all
3. What cognitive biases feel like from the inside
4. Coordination as a Scarce Resource
5. 2018 Review: Voting Results!
6. Moral public goods
7. On hiding the source of knowledge
8. AI Alignment 2018-19 Review
9. Becoming Unusually Truth-Oriented
10. Hedonic asymmetries
11. Technology Changes Constraints
12. Of arguments and wagers
13. A rant against robots
14. Review: How to Read a Book (Mortimer Adler, Charles Van Doren)
15. Draining the swamp
16. Exploring safe exploration
17. Why Do You Keep Having This Problem?
18. The Road to Mazedom
19. Circling as Cousin to Rationality
20. Don't Double-Crux With Suicide Rock
21. Criticism as Entertainment
22. The Alignment-Competence Trade-Off, Part 1: Coalition Size and Signaling Costs
23. Studying Early Stage Science: Research Program Introduction
24. Homeostasis and “Root Causes” in Aging
25. What we Know vs. How we Know it?
26. Does GPT-2 Understand Anything?
27. Book review: Rethinking Consciousness
28. Material Goods as an Abundant Resource
29. Clarifying The Malignity of the Universal Prior: The Lexical Update
30. How Doomed are Large Organizations?
31. Appendix: how a subagent could get powerful
32. Use-cases for computations, other than running them?
33. Excitement vs childishness
34. How to Escape From Immoral Mazes
35. Subscripting Typographic Convention For Citations/Dates/Sources/Evidentials: A Proposal
36. The two-layer model of human values, and problems with synthesizing preferences
37. Disasters
38. [AN #83]: Sample-efficient deep learning with ReMixMatch
39. How to Identify an Immoral Maze
40. Moral uncertainty: What kind of 'should' is involved?
41. Healing vs. exercise analogies for emotional work
42. Using Vickrey auctions as a price discovery mechanism
43. Normalization of Deviance
44. Why a New Rationalization Sequence?
45. Update on Ought's experiments on factored evaluation of arguments
46. Safety regulators: A tool for mitigating technological risk

47. [AN #81]: Universality as a potential solution to conceptual difficulties in intent alignment
48. UML VII: Meta-Learning
49. Algorithms vs Compute
50. What is Life in an Immoral Maze?

Reality-Revealing and Reality-Masking Puzzles

Tl;dr: I'll try here to show how CFAR's "art of rationality" has evolved over time, and what has driven that evolution.

In the course of this, I'll introduce the distinction between what I'll call "reality-revealing puzzles" and "reality-masking puzzles"—a distinction that I think is almost necessary for anyone attempting to develop a psychological art in ways that will help rather than harm. (And one I wish I'd had explicitly back when the Center for Applied Rationality was founded.)

I'll also be trying to elaborate, here, on the notion we at CFAR have recently been tossing around about CFAR being an attempt to bridge between common sense and Singularity scenarios—an attempt to figure out how people can stay grounded in common sense and ordinary decency and humane values and so on, while also taking in (and planning actions within) the kind of universe we may actually be living in.

--

Arts grow from puzzles. I like to look at mathematics, or music, or ungodly things like marketing, and ask: What puzzles were its creators tinkering with that led them to leave behind these structures? (Structures now being used by other people, for other reasons.)

I picture arts like coral reefs. Coral polyps build shell-bits for their own reasons, but over time there accumulates a reef usable by others. Math built up like this—and math is now a powerful structure for building from. [Sales and Freud and modern marketing/self-help/sales etc. built up some patterns too—and our basic way of seeing each other and ourselves is now built partly in and from all these structures, for better and for worse.]

So let's ask: What sort of reef is CFAR living within, and adding to? From what puzzles (what patterns of tinkering) has our "rationality" accumulated?

Two kinds of puzzles: "reality-revealing" and "reality-masking"

First, some background. Some puzzles invite a kind of tinkering that lets the world in and leaves you smarter. A kid whittling with a pocket knife is entangling her mind with bits of reality. So is a driver who notices something small about how pedestrians dart into streets, and adjusts accordingly. So also is the mathematician at her daily work. And so on.

Other puzzles (or other contexts) invite a kind of tinkering that has the opposite effect. They invite a tinkering that gradually figures out how to mask parts of the world from your vision. For example, some months into my work as a math tutor I realized I'd been unconsciously learning how to cue my students into acting like my words made sense (even when they didn't). I'd learned to mask from my own senses the clues about what my students were and were not learning.

We'll be referring to these puzzle-types a lot, so it'll help to have a term for them. I'll call these puzzles "good" or "reality-revealing" puzzles, and "bad" or "reality-masking" puzzles, respectively. Both puzzle-types appear abundantly in most folks' lives, often mixed together. The same kid with the pocket knife who is busy entangling her mind with data about bark and woodchips and fine motor patterns (from the "good" puzzle of "how can I whittle this stick"), may simultaneously be busy tinkering with the "bad" puzzle of "how can I not-notice when my creations fall short of my hopes."

(Even "good" puzzles can cause skill loss: a person who studies Dvorak may lose some of their QWERTY skill, and someone who adapts to the unselfconscious arguing of the math department may do worse for a while in contexts requiring tact. The distinction is that "good" puzzles do this *only incidentally.* Good puzzles do not invite a *search for* configurations that mask bits of reality. Whereas with me and my math tutees, say, there was a direct reward/conditioning response that happened specifically when the "they didn't get it" signal was masked from my view. There was a small optimizer inside of me that was *learning* how to mask parts of the world from me, *via feedback from the systems of mine it was learning to befuddle*.)

Also, certain good puzzles (and certain bad ones!) allow unusually powerful accumulations across time. I'd list math, computer science, and the English language as examples of unusually powerful artifacts for improving vision. I'd list "sales and marketing skill" as an example of an unusually powerful artifact for impairing vision (the salesperson's own vision, not just the customer's).

The puzzles that helped build CFAR

Much of what I love about CFAR is linked to the puzzles we dwell near (the reality-revealing ones, I mean). And much of what gives me the shudders about CFAR comes from a reality-masking puzzle-set that's been interlinked with these.

Eliezer created the Sequences after staring a lot at the AI alignment problem. He asked how a computer system could form a "map" that matches the territory; he asked how he himself could do the same. He asked, "Why do I believe what I believe?" and checked whether the *mechanistic causal history* that gave rise to his beliefs would have yielded *different beliefs in a world where different things were true.*

There's a springing up into self-awareness that can come from this! A taking hold of our power as humans to see. A child's visceral sense that of course we care and should care—freed from its learned hopelessness. And taking on the stars themselves with daring!

CFAR took these origins and worked to make at least parts of them accessible to some who bounced off the Sequences, or who wouldn't have read the Sequences. We created feedback loops for practicing some of the core Sequences-bits in the context of folks' ordinary lives rather than in the context of philosophy puzzles. If you take a person (even a rather good scientist) and introduce them to the questions about AI and the long-term future… often nothing much happens in their head except some random stuck nonsense intuitions ("AIs wouldn't do that, because they're our offspring. What's for lunch?"). So we built a way to practice some of the core moves that alignment thinking needed. Especially, we built a way to practice *having thoughts at all, in cases where standard just-do-what-the-neighbors-do strategies would tend to block them off.*

For example:

- *Inner Simulator.* (Your "beliefs" are what you expect to see happen—not what you "endorse" on a verbal level. You can practice tracking these anticipations in daily life! And making plans with them! And once you've seen that they're useful for planning—well, you might try also having them in contexts like AI risk. Turns out you have beliefs even where you don't have official "expertise" or credentials authorizing belief-creation! And you can dialog with them, and there's sense there.)
- *Crux-Mapping; Double Crux.* (Extends your ability to dialog with inner simulator-style beliefs. Lets you find in yourself a random opaque intuition about AI being [likely/unlikely/safe/whatever], and then query it via thought experiments until it is more made out of introspectable verbal reasoning. Lets two people with different intuitions collide them in verbal conversation.)
- *Goal Factoring and Units of Exchange.* (Life isn't multiple choice; you can name the good things and the bad things, and you can invest in seeking the alternatives with more of the good and less of the bad. For example, if you could save 4 months in a world where you were allowed to complete your PhD early, it may be worth more than several hours to scheme out how to somehow purchase permission from your advisor, since 4 months is worth rather more than several hours.)
- *Hamming Questions.* (Some questions are worth a lot more than others. You want to focus at least some of your attention on the most important questions affecting your life, rather than just the random details in front of you. And you can just decide to do that on purpose, by using pen and paper and a timer!)[1]

Much good resulted from this—many loved the Sequences; many loved CFAR's intro workshops; and a fair number who started there went into careers in AI alignment work and credited CFAR workshops as partially causal.

And still, as we did this, problems arose. AI risk is disorienting! Helping AI risk hit more people meant "helping" more people encounter something disorienting. And so we set to work on that as well. The thing I would say now about the reality-revealing puzzles that helped grow CFAR is that there were three, each closely linked with each other:

1. Will AI at some point radically transform our lightcone? (How / why / with what details and intervention options?)
2. How do we get our minds to make contact with problem (1)? And how do we think groundedly about such things, rather than having accidental nonsense-intuitions and sticking there?
3. How do we stay human, and stay reliably in contact with what's worth caring about (valuing honesty and compassion and hard work; having reliable friendships; being good people and good thinkers and doers), while still taking in how disorientingly different the future might be? (And while neither pretending that we have no shot at changing the future, nor that "what actions should I take to impact the future?" is a multiple choice question with nothing further to do, nor that any particular silly plan is more likely to work than it is?)

CFAR grew up around all three of these puzzles—but (2) played an especially large role over most of our history, and (3) has played an especially large role over the last year and (I think) will over the coming one.

I'd like to talk now about (3), and about the disorientation patterns that make (3) needed.

Disorientation patterns

To start with an analogous event: The process of losing a deeply held childhood religion can be quite disruptive to a person's common sense and values. Let us take as examples the two commonsensical statements:

- (A) It is worth getting out of bed in the morning; and,
- (B) It is okay to care about my friends.

These two commonsensical statements are held by most religious people. They are actually also held by most atheists. Nevertheless, when a person loses their religion, they fairly often become temporarily unsure about whether these two statements (and various similar such statements) are true. That's because somehow the person's understanding of why statements (A) and (B) are true was often tangled up in (for example) Jehovah. And figuring out how to think about these things in the absence of their childhood religion (even in cases like this one where the statements should survive!) can require actual work. (This is particularly true because some things really are different given that Jehovah is false—and it can take work to determine which is which.)

Over the last 12 years, I've chatted with small hundreds of people who were somewhere "in process" along the path toward "okay I guess I should take Singularity scenarios seriously." From watching them, my guess is that the process of coming to take Singularity scenarios seriously is often even more disruptive than is losing a childhood religion. Among many other things, I have seen it sometimes disrupt:

- People's belief that they should have rest, free time, some money/time/energy to spend on objects of their choosing, abundant sleep, etc.
    - "It used to be okay to buy myself hot cocoa from time to time, because there used to be nothing important I could do with money. But now—should I never buy hot cocoa? Should I agonize freshly each time? If I do buy a hot cocoa does that mean I don't care?"
- People's in-practice ability to "hang out"—to enjoy their friends, or the beach, in a "just being in the moment" kind of way.
    - "Here I am at the beach like my to-do list told me to be, since I'm a good EA who is planning not to burn out. I've got my friends, beer, guitar, waves: check. But how is it that I used to be able to enter "hanging out mode"? And why do my friends keep making meaningless mouth-noises that have nothing to do with what's eventually going to happen to everyone?"
- People's understanding of whether commonsense morality holds, and of whether they can expect other folks in this space to also believe that commonsense morality holds.
    - "Given the vast cosmic stakes, surely doing the thing that is expedient is more important than, say, honesty?"
- People's in-practice tendency to have serious hobbies and to take a deep interest in how the world works.
    - "I used to enjoy learning mathematics just for the sake of it, and trying to understand history for fun. But it's actually jillions of times higher value to work on [decision theory, or ML, or whatever else is pre-labeled as 'AI risk relevant']."
- People's ability to link in with ordinary institutions and take them seriously (e.g. to continue learning from their day job and caring about their colleagues'

progress and problems; to continue enjoying the dance club they used to dance at; to continue to take an interest in their significant other's life and work; to continue learning from their PhD program; etc.)
    - "Here I am at my day job, meaninglessly doing nothing to help no one, while the world is at stake—how is it that before learning about the Singularity, I used to be learning skills and finding meaning and enjoying myself in this role?"
- People's understanding of what's worth caring about, or what's worth fighting for
    - "So… 'happiness' is valuable, which means that I should hope we get an AI that tiles the universe with a single repeating mouse orgasm, right? ... I wonder why imagining a 'valuable' future doesn't feel that good/motivating to me."
- People's understanding of when to use their own judgment and when to defer to others.
    - "AI risk is really really important… which probably means I should pick some random person at MIRI or CEA or somewhere and assume they know more than I do about my own career and future, right?"

My take is that many of these disorientation-bits are analogous to the new atheist's disorientation discussed earlier. "Getting out of bed in the morning" and "caring about one's friends" turn out to be useful for more reasons than Jehovah—but their derivation in the mind of that person was entangled with Jehovah. Honesty is analogously valuable for more reasons than its value as a local consumption good; and many of these reasons apply extra if the stakes are high. But the derivation of honesty that many folks were raised with does not survive the change in imagined surroundings—and so it needs to be thought through freshly.

Another part of the disorientation perhaps stems from emotional reeling in contact with the possibility of death (both one's own death, and the death of the larger culture/tribe/species/values/life one has been part of).

And yet another part seems to me to stem from a set of "bad" puzzles that were inadvertently joined with the "good" puzzles involved in thinking through Singularity scenarios—"bad" puzzles that disable the mental immune systems that normally prevent updating in huge ways from weird and out-there claims. I'll postpone this third part for a section and then return to it.

There is value in helping people with this disorientation; and much of this helping work is tractable

It seems not-surprising that people are disrupted in cases where they seriously, viscerally wonder "Hey, is everything I know and everything humanity has ever been doing to maybe-end, and also to maybe become any number of unimaginably awesome things? Also, am I personally in a position of possibly incredibly high leverage and yet also very high ambiguity with respect to all that?"

Perhaps it is more surprising that people in fact sometimes let this into their system 1's at all. Many do, though; including many (but certainly not all!) of those I would consider highly effective. At least, I've had many many conversations with people who seem viscerally affected by all this. Also, many people who tell me AI risk is "only abstract to [them]" still burst into tears or otherwise exhibit unambiguous strong

emotion when asked certain questions—so I think people are sometimes more affected than they think.

An additional point is that many folks over the years have told me that they were choosing not to think much about Singularity scenarios lest such thinking destabilize them in various ways. I suspect that many who are in principle capable of doing useful technical work on AI alignment presently avoid the topic for such reasons. Also, many such folks have told me over the years that they found pieces at CFAR that allowed them to feel more confident in attempting such thinking, and that finding these pieces then caused them to go forth and attempt such thinking. (Alas, I know of at least one person who later reported that they had been *inaccurate* in revising this risk assessment! Caution seems recommended.)

Finally: people sometimes suggest to me that researchers could dodge this whole set of difficulties by simply reasoning about Singularity scenarios abstractly, while avoiding ever letting such scenarios get into their viscera. While I expect such attempts are in fact useful to some, I believe this method insufficient for two reasons. First, as noted, it seems to me that these topics sometimes get under people's skin more than they intend or realize. Second, it seems to me that visceral engagement with the AI alignment problem is often helpful for the best scientific research—if a person is to work with a given "puzzle" it is easier to do so when they can concretely picture the puzzle, including in their system 1. This is why mathematicians often take pains to "understand why a given theorem is true" rather than only to follow its derivation abstractly. This is why Richard Feynman took pains to picture the physics he was working with in the "make your beliefs pay rent in anticipated experiences" sense and took pains to ensure that his students could link phrases such as "materials with an index of refraction" with examples such as "water." I would guess that with AI alignment research, as elsewhere, it is easier to do first-rate scientific work when you have visceral models of what the terms, claims, and puzzles mean and how it all fits together.

In terms of the tractability of assisting with disorientation in such cases: it seems to me that simply providing contexts for people to talk to folks who've "been there before" can be pretty helpful. I believe various other concepts we have are also helpful, such as: familiarity with what bucket errors often look like for AI risk newcomers; discussion of the unilateralist's curse; explanations of why hobbies and world-modeling and honesty still matter when the stakes are high. (Certainly participants sometimes say that these are helpful.) The assistance is partial, but there's a decent iteration loop for tinkering away at it. We'll also be trying some LessWrong posts on some of this in the coming year.

A cluster of "reality-masking" puzzles that also shaped CFAR

To what extent has CFAR's art been shaped by reality-masking puzzles—tinkering loops that inadvertently disable parts of our ability to see? And how can we tell, and how can we reduce such loops? And what role have reality-masking puzzles played in the disruption that sometimes happens to folks who get into AI risk (in and out of CFAR)?

My guess is actually that a fair bit of this sort of reality-masking has occurred. (My guess is that the amount is "strategically significant" but not "utterly overwhelming.")

To name one of the more important dynamics:

Disabling pieces of the epistemic immune system

Folks arrive with piles of heuristics that help them avoid nonsense beliefs and rash actions. Unfortunately, many of these heuristics—including many of the generally useful ones—can "get in the way." They "get in the way" of thinking about AI risk. They also "get in the way" of folks at mainline workshops thinking about changing jobs/relationships/life patterns etc. unrelated to AI risk. And so disabling them can sometimes help people acquire accurate beliefs about important things, and have more felt freedom to change their lives in ways they want.

Thus, the naive process of tinkering toward "really helping this person think about AI risk" (or "really helping this person consider their life options and make choices") can lead to folks disabling parts of their epistemic immune system. (And unfortunately also thereby disabling their future ability to detect certain classes of false claims!)

For example, the Sequences make some effort to disable:

- The absurdity heuristic
- The habit of saying/believing one "doesn't know" in cases where one hasn't much "legitimate" evidence
- Compartmentalization

Similarly, CFAR workshops sometimes have the effect of disabling:

- Taste as a fixed guide to which people/organizations/ideas to take in or to spit out. (People come in believing that certain things just "are" yucky. Then, we teach them how to "dialog" with their tastes… and they become more apt to sometimes-ignore previous "yuck" reactions.)
- Antibodies that protect people from updating toward optimizing for a specific goal, rather than for a portfolio of goals. For example, entering participants will say things like "I know it's not rational, but I also like to [activity straw vulcans undervalue]." And even though CFAR workshops explicitly warn against straw vulcanism, they also explicitly encourage people to work toward having goals that are more internally consistent, which sometimes has the effect of disabling the antibody which prevents people from suddenly re-conceptualizing most of their goal set as all being instrumental to/in service of some particular purportedly-paramount goal.
- Folks' tendency to take actions based on social roles (e.g., CFAR's Goal-Factoring class used to explicitly teach people not to say "I'm studying for my exam because I'm a college student" or "I have to do it because it's my job," and to instead say "I'm studying for my exam in order to [cause outcome X]").

Again, these particular shifts are not all bad; many of them have advantages. But I think their costs are easy to underestimate, and I'm interested in seeing whether we can get a "rationality" that causes less disablement of ordinary human patterns of functioning, while still helping people reason well in contexts where there aren't good prexisting epistemic guardrails. CFAR seems likely to spend a good bit of time modeling these problems over the coming year, and trying to develop candidate solutions—we're already playing with a bunch of new curriculum designed primarily for this purpose—and we'd love to get LessWrong's thoughts before playing further!

Acknowledgement

Thanks to Adam Scholl for helping a lot with the writing. Remaining flaws are of course my own.

Edited to add:

I think I did not spell out well enough what I mean by "reality-masking puzzles." I try again in a comment.

I think that getting this ontology right is a core and difficult task, and one I haven't finished solving yet -- it is the task of finding analogs of the "reasoning vs rationalization" distinction that are suitable for understanding group dynamics. I would love help with this task -- that is maybe the main reason I wrote this post.

I think this task is closely related to what Zvi and the book "Moral Mazes" are trying for.

---

1. If you don't know some of these terms but want to, you can find them in CFAR's handbook. ↩

CFAR Participant Handbook now available to all

Google Drive PDF

Hey, guys—I wrote this, and CFAR has recently decided to make it publicly available. Much of it involved rewriting the original work of others, such as Anna Salamon, Kenzie Ashkie, Val Smith, Dan Keys, and other influential CFAR founders and staff, but the actual content was filtered through me as single author as part of getting everything into a consistent and coherent shape.

I have mild intentions to update it in the future with a handful of other new chapters that were on the list, but which didn't get written before CFAR let me go. Note that such updates will likely not be current-CFAR-approved, but will still derive directly from my understanding of the curriculum as former Curriculum Director.

What cognitive biases feel like from the inside

Building on the recent SSC post [Why Doctors Think They're The Best](...)...

What it feels like for me	How I see others who feel the same
There is controversy on the subject but there shouldn't be because the side I am on is obviously right.	They have taken one side in a debate that is unresolved for good reason that they are struggling to understand
I have been studying this carefully	They preferentially seek out conforming evidence
The arguments for my side make obvious sense, they're almost boring.	They're very ready to accept any and all arguments for their side.
The arguments for the opposing side are contradictory, superficial, illogical or debunked.	They dismiss arguments for the opposing side at the earliest opportunity.
The people on the opposing side believe these arguments mostly because they are uninformed, have not thought about it enough or are being actively misled by people with bad motives.	The flawed way they perceive the opposing side makes them confused about how anyone could be on that side. They resolve that confusion by making strong assumptions that can approach conspiracy theories.

The scientific term for this mismatch is: **confirmation bias**

What it feels like for me	How I see others who feel the same
My customers/friends/relationships love me, so I am good for them, so I am probably just generally good.	They neglect the customers / friends / relationships that did not love them and have left, so they overestimate how good they are.
When customers / friends / relationships switch to me, they tell horror stories of who I'm replacing for them, so I'm better than those.	They don't see the people who are happy with who they have and therefore never become their customers / friends / relationships.

The scientific term for this mismatch is: **selection bias**

What it feels like for me	How I see others who feel the same
Although I am smart and friendly, people don't listen to me.	Although they are smart and friendly, they are hard to understand.
I have a deep understanding of the issue that people are too	They are failing to communicate their understanding, or to give unambiguous evidence they even have it.

stupid or too disinterested to come to share.	
This lack of being listened to affects several areas of my life but it is particularly jarring on topics that are very important to me.	This bad communication affects all areas of their life, but on the unimportant ones they don't even understand that others don't understand them.

The scientific term for this mismatch is: **illusion of transparency**

What it feels like for me	**How I see others who feel the same**
I knew at the time this would not go as planned.	They did not predict what was going to happen.
The plan was bad and we should have known it was bad.	They fail to appreciate how hard prediction is, so the mistake seems more obvious to them than it was.
I knew it was bad, I just didn't say it, for good reasons (e.g. out of politeness or too much trust in those who made the bad plan) or because it is not my responsibility or because nobody listens to me anyway.	In order to avoid blame for the seemingly obvious mistake, they are making up excuses.

The scientific term for this mismatch is: **hindsight bias**

What it feels like for me	**How I see others who feel the same**
I have a good intuition; even decisions I make based on insufficient information tend to turn out to be right.	They tend to recall their own successes and forget their own failures, leading to an inflated sense of past success.
I know early on how well certain projects are going to go or how well I will get along with certain people.	They make self-fulfilling prophecies that directly influence how much effort they put into a project or relationship.
Compared to others, I am unusually successful in my decisions.	They evaluate the decisions of others more level-headedly than their own.
I am therefore comfortable relying on my quick decisions.	They therefore overestimate the quality of their decisions.
This is more true for life decisions that are very important to me.	Yes, this is more true for life decisions that are very important to them.

The scientific term for this mismatch is: **optimism bias**

Why this is better than how we usually talk about biases

Communication in abstracts is very hard. (See: Illusion of Transparency: Why No One Understands You) Therefore, it often fails. (See: Explainers Shoot High. Aim Low!) It is hard to even notice communication has failed. (See: Double Illusion of Transparency) Therefore it is hard to appreciate how rarely communication in abstracts actually succeeds.

Rationalists have noticed this. (Example) Scott Alexander uses a lot of concrete examples and that should be a major reason why he's our best communicator. Eliezer's Sequences work partly because he uses examples and even fiction to illustrate. But when the rest of us talk about rationality we still mostly talk in abstracts.

For example, this recent video was praised by many for being comparatively approachable. And it does do many things right, such as emphasize and repeat that evidence alone should not generate probabilities, but should only ever update prior probabilities. But it still spends more than half of its runtime displaying mathematical notation that no more than 3% of the population can even read. For the vast majority of people, only the example it uses can possibly "stick". Yet the video uses its single example as no more than a means for getting to the abstract explanation.

This is a mistake. I believe a video with three to five vivid examples of how to apply Bayes' Theorem, preferably funny or sexy ones, would leave a much more lasting impression on most people.

Our highly demanding style of communication correctly predicts that LessWrongians are, on average, much smarter, much more STEM-educated and much younger than the general population. You have to be that way to even be able to drink the Kool Aid! This makes us homogeneous, which is probably a big part of what makes LW feel tribal, which is emotionally satisfying. But it leaves most of the world with their bad decisions. We need to be Raising the Sanity Waterline and we can't do that by continuing to communicate largely in abstracts.

The tables above show one way to do better that does the following.

- It aims low - merely to help people notice the flaws in their thinking. It will not, and does not need to, enable readers to write scientific papers on the subject.
- It reduces biases into mismatches between Inside View and Outside View. It lists concrete observations from both views and juxtaposes them.
- These observations are written in a way that is hopefully general enough for most people to find they match their own experiences.
- It trusts readers to infer from these juxtaposed observations their own understanding of the phenomena. After all, generalizing over particulars is much easier than integrating generalizations and applying them to particulars. The understanding gained this way will be imprecise, but it has the advantage of actually arriving inside the reader's mind.
- It is nearly jargon free; it only names the biases for the benefit of that small minority who might want to learn more.

What do you think about this? Should we communicate more concretely? If so, should we do it in this way or what would you do differently?

Would you like to correct these tables? Would you like to propose more analogous observations or other biases?

Thanks to Simon, miniBill and others for helping with the draft of this post.

Coordination as a Scarce Resource

Let’s start with a few examples of very common real-world coordination problems.

- The marketing department at a car dealership posts ads for specific cars, but the salespeople don’t know which cars were advertised, causing confusion when a customer calls in asking about a specific car. There’s no intentional information-hoarding, it’s just that the marketing and sales people don’t sit next to each other or talk very often. Even if the info were shared, it would need to be translated to a format usable by the salespeople.
- Various hard problems in analysis of large-scale biological data likely have close analogues in econometrics. The econometricians have good methods to solve the problems, and would probably be quite happy to apply those methods to biological data, and the bio experimentalists would love some analytic help. But these people hardly ever talk to each other, and use different language for the same things anyway.
- When the US invaded Grenada in the ‘80’s, the marines occupied one side of the island and the army occupied the other. Their radios were not compatible, so if an army officer needed to contact their counterpart in the marines, they had to walk to the nearest pay phone and get routed through Fort Bragg on commercial telephone lines.
- Various US intelligence agencies had all of the pieces necessary to stop the 9/11 attacks. There were agencies which knew something was planned for that day, and knew who the actors were. There were agencies which knew the terrorists were getting on the planes. There were agencies which could have moved to stop them, but unfortunately the fax(!) from the agencies which knew what was happening wasn’t checked in time.
- There are about 300 million people in the US. If I have a small company producing doilies, chances are there are plenty of people in the US alone who’d love my doilies and be happy to pay for them. But it’s hard to figure out exactly which people those are, and even once that’s done it’s hard to get them a message showing off my product. And even if all that works out, if the customers really want a slightly different pattern, it’s hard for them to communicate back to me what they want - even if I’d be happy to make it.

So coordination problems are a constraint to production of all kinds of economic value. How taut are those constraints?

Well, let’s look at the market price of relaxing coordination constraints. In other words: how much do people/companies get paid for solving coordination problems?

When I think of people whose *main* job is to solve coordination problems, here are some occupations which spring to mind:

- Entrepreneurs’ main job is to coordinate salespeople, engineers, designers, marketers, investors, customers, regulators, suppliers, shippers, etc…
- Managers’ main job is to coordinate between their bosses, underlings, and across departments
- Investment bankers coordinate between investors, companies, lawyers, and a huge number of people within each of those organizations
- Real estate developers coordinate between builders, landowners, regulators, renters, and investors

Note that all of these are occupations typically associated with very high pay. Even more to the point: within each of these occupations, people who solve more complicated coordination problems (e.g. between more people) tend to make more money. Even at the small end, the main difference between an employee and a freelancer is that the freelancer has to solve their own coordination problem (i.e. find people who want their services); freelancers make lots of money mainly when they are very good at solving this problem.

Similarly with companies. If we go down the list of tech unicorns, most (though not all) of them solve coordination problems as their primary business model:

- Google matches company websites to potentially interested users
- Facebook is a general-purpose coordination platform
- Amazon and Ebay are general-purpose marketplaces: they match buyers to sellers
- Uber/Lyft are more specialized marketplaces

Again: solving coordination problems at scale offers huge amounts of money.

This suggests that coordination problems are *very* taut constraints in today's economy.

It's not hard to imagine *why* coordination problems would be very taut today. Over the past ~50 years, global travel/transportation has gone from rare to ordinary, and global communication has become cheap and ubiquitous. Geographical constraints have largely been relaxed, and communication/information processing constraints have largely been relaxed. The number of people we *could* potentially coordinate with has expanded massively as a result - a small doily business can now sell to a national or even global customer base; a phone app can connect any willing driver in a city to any paying rider.

Yet human brains have not changed much, even as the number of people we interact with skyrockets past Dunbar's number. It's hard for humans to coordinate with thousands - let alone billions - of other humans. The coordination constraint remains, so as other constraints relax, it becomes more taut.

What Would This Model Predict?

One prediction: suppose I've decided to become a freelancer/consultant. I can invest effort in becoming better at my object-level craft, or I can invest effort in becoming better at solving my coordination problem - e.g. by exploring marketing channels or studying my target market. Which of these will make more money? Probably the latter - coordination constraints are *very* taut, so there's lots of money to be made by relaxing them.

More generally, when evaluating new business ideas, questions on my short list include:

- How will this business find people who would want to buy the product?
- How will this business make those people aware of the product?

To the extent that coordination is an unusually taut constraint, answers to these questions will be the main determinant of business profitability - even more so than product quality.

Coming from a different direction, when considering a business idea we should ask how many different kinds of people this business needs to coordinate. At one point I worked at a mortgage startup, where the list of internal departments included marketing, sales, underwriting, legal, and capital markets on the mortgage side, plus design, engineering, and ops on the tech side, and on top of that we had to interface to at least a dozen external companies on a regular basis. Coordination is *the* primary constraint at a company like that.

Yet another direction: if coordination constraints are very taut, then we expect adoption of technology which makes coordination easier. One form of this is outlined in From Personal to Prison Gangs: people make *themselves* easier to coordinate with, by following standardized patterns of behavior and fitting into standard molds. For instance, in large organizations (where more people need to coordinate) we tend to see group-based identity: rather than understanding what John or Allan does, people understand what lawyers or developers do. Interactions between people become more standardized, and roles more rigid - these are solutions to coordination problems. Such solutions entail large tradeoffs, but coordination constraints are very taut, so large tradeoffs are accepted.

Conversely, if we want a world with less pressure to standardize behavior, then we need some other way to relax coordination constraints - some technology which helps people coordinate at scale without needing to standardize behavior as much. Such technology would probably see wide adoption, and potentially make quite a lot of money as well.

Summary

I often hear people they'd like object-level skill/effort to be rewarded more than marketing/sales, or they'd like to see less pressure to standardize behavior, or they'd like the world to be more individualized and identity to be less group-based. To the extent that we buy the picture here, all of these phenomena are solutions to coordination problems. Society rewards marketing over object-level skill, and tries to standardize behavior, because coordination constraints are extremely taut.

If we want the world to look less like that, then we need alternative scalable technologies to solve coordination problems.

2018 Review: Voting Results!

The votes are in!

59 of the 430 eligible voters participated, evaluating 75 posts. Meanwhile, 39 users submitted a total of 120 reviews, with most posts getting at least one review.

Thanks a ton to everyone who put in time to think about the posts - nominators, reviewers and voters alike. Several reviews substantially changed my mind about many topics and ideas, and I was quite grateful for the authors participating in the process. I'll mention Zack_M_Davis, Vanessa Kosoy, and Daniel Filan as great people who wrote the most upvoted reviews.

In the coming months, the LessWrong team will write further analyses of the vote data, and use the information to form a sequence and a book of the best writing on LessWrong from 2018.

Below are the results of the vote, followed by a discussion of how reliable the result is and plans for the future.

Top 15 posts

1. Embedded Agents by Abram Demski and Scott Garrabrant
2. The Rocket Alignment Problem by Eliezer Yudkowsky
3. Local Validity as a Key to Sanity and Civilization by Eliezer Yudkowsky
4. Arguments about fast takeoff by Paul Christiano
5. The Costly Coordination Mechanism of Common Knowledge by Ben Pace
6. Toward a New Technical Explanation of Technical Explanation by Abram Demski
7. Anti-social Punishment by Martin Sustrik
8. The Tails Coming Apart As Metaphor For Life by Scott Alexander
9. Babble by alkjash
10. The Loudest Alarm Is Probably False by orthonormal
11. The Intelligent Social Web by Valentine
12. Prediction Markets: When Do They Work? by Zvi
13. Coherence arguments do not imply goal-directed behavior by Rohin Shah
14. Is Science Slowing Down? by Scott Alexander
15. A voting theory primer for rationalists by Jameson Quinn and Robustness to Scale by Scott Garrabrant

Top 15 posts not about AI

1. Local Validity as a Key to Sanity and Civilization by Eliezer Yudkowsky
2. The Costly Coordination Mechanism of Common Knowledge by Ben Pace
3. Anti-social Punishment by Martin Sustrik
4. The Tails Coming Apart As Metaphor For Life by Scott Alexander
5. Babble by alkjash
6. The Loudest Alarm Is Probably False by Orthonormal
7. The Intelligent Social Web by Valentine
8. Prediction Markets: When Do They Work? by Zvi
9. Is Science Slowing Down? by Scott Alexander
10. A voting theory primer for rationalists by Jameson Quinn
11. Toolbox-thinking and Law-thinking by Eliezer Yudkowsky
12. A Sketch of Good Communication by Ben Pace
13. A LessWrong Crypto Autopsy by Scott Alexander
14. Unrolling social metacognition: Three levels of meta are not enough. by Academian
15. Varieties Of Argumentative Experience by Scott Alexander

Top 10 posts about AI

(The vote included 20 posts about AI.)

1. Embedded Agents by Abram Demski and Scott Garrabrant
2. The Rocket Alignment Problem by Eliezer Yudkowsky
3. Arguments about fast takeoff by Paul Christiano
4. Toward a New Technical Explanation of Technical Explanation by Abram Demski
5. Coherence arguments do not imply goal-directed behavior by Rohin Shah
6. Robustness to Scale by Scott Garrabrant
7. Paul's research agenda FAQ by zhukeepa
8. An Untrollable Mathematician Illustrated by Abram Demski
9. Specification gaming examples in AI by Vika
10. 2018 AI Alignment Literature Review and Charity Comparison by Larks

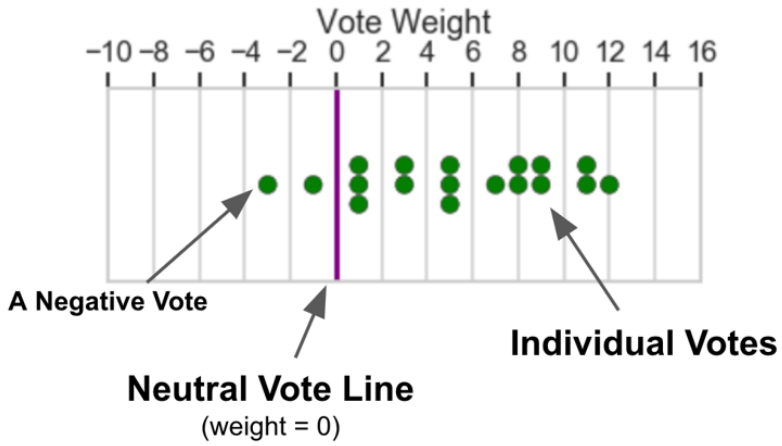
The Complete Results

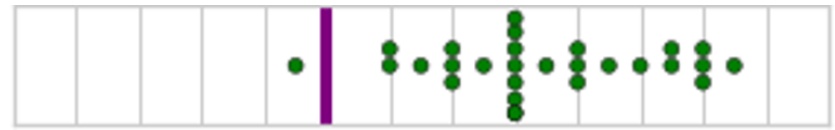
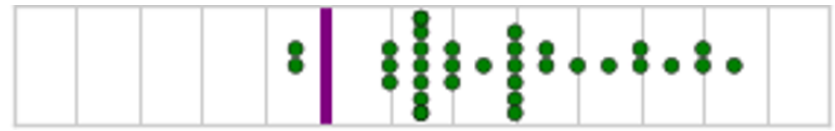
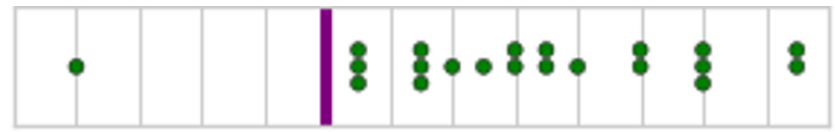
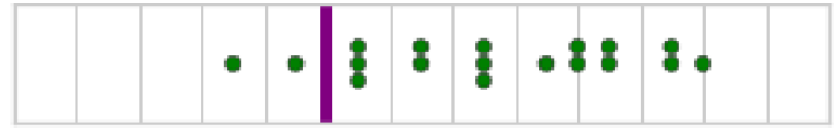
Click Here If You Would Like A More Comprehensive Vote Data Spreadsheet

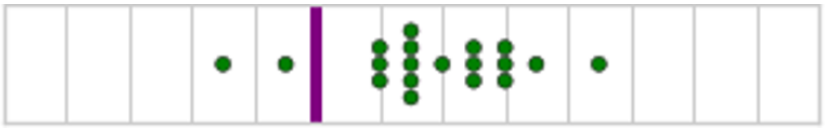
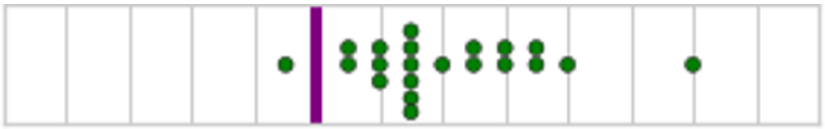
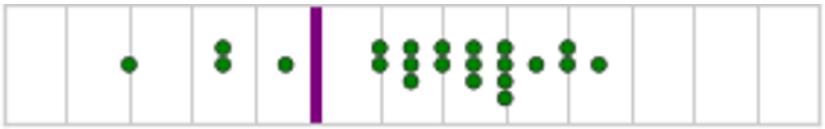
To help users see the spread of the vote data, we've included *swarmplot* visualizations.

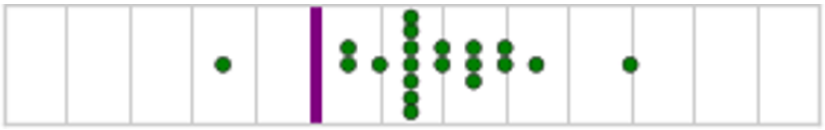
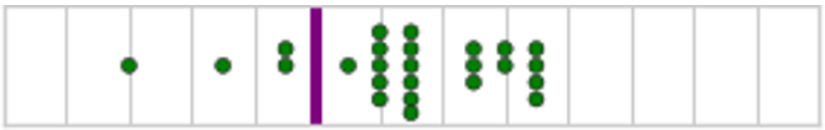
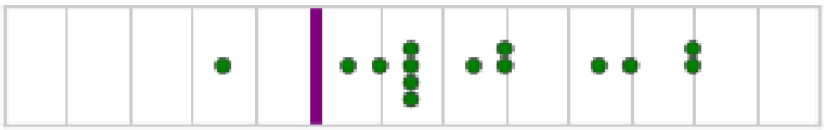
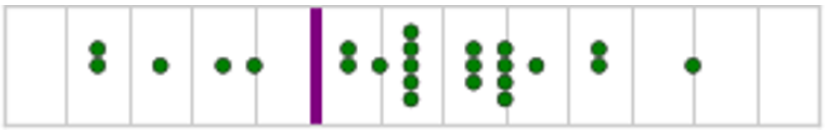
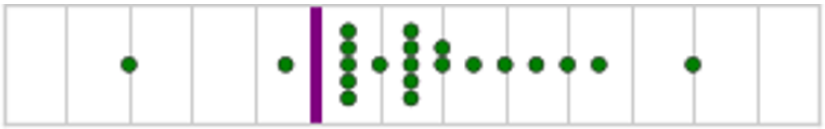
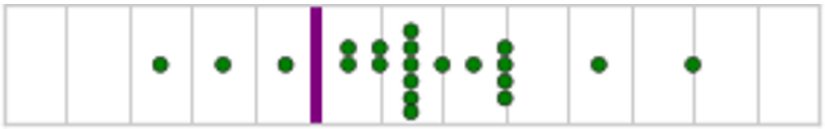
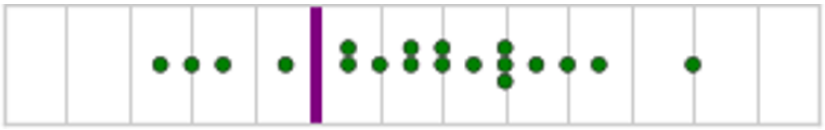
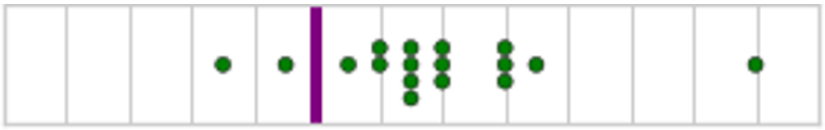
- For space reasons, only votes with weights between -10 and 16 are plotted. This covers 99.4% of votes.
- Gridlines are spaced 2 points apart.

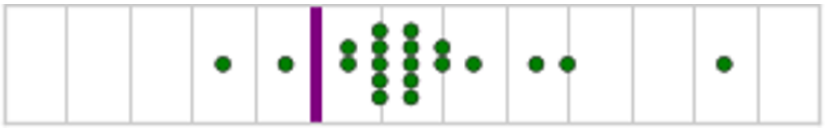
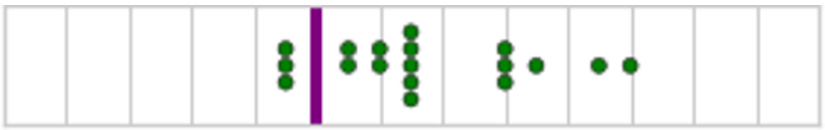
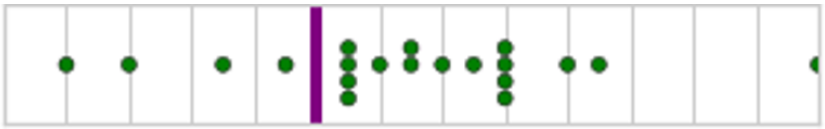
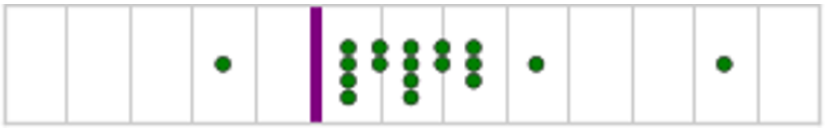
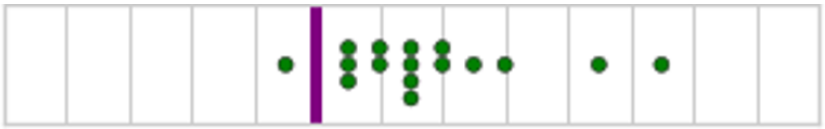
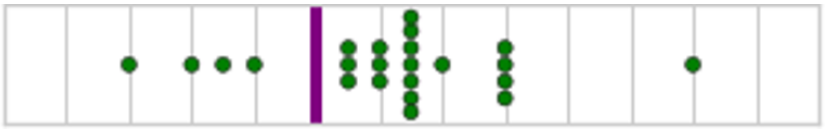
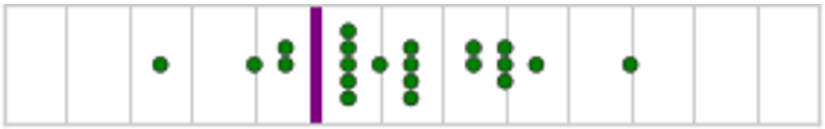
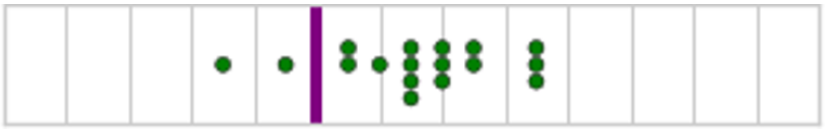
- Concrete illustration: The plot immediately below has 18 votes ranging in strength from -3 to 12.

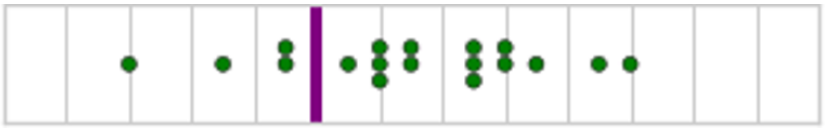
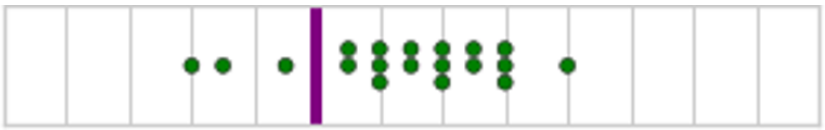
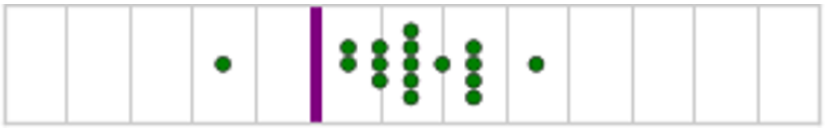
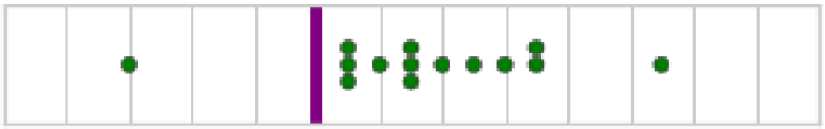
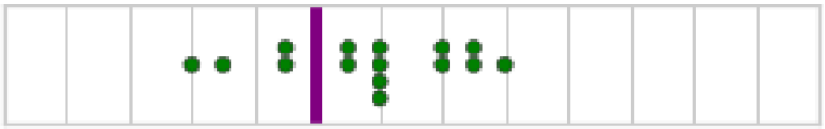
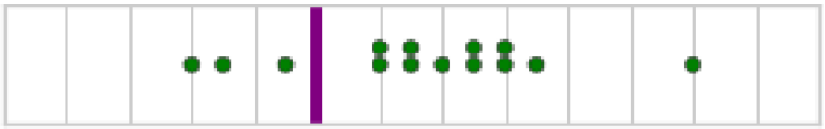
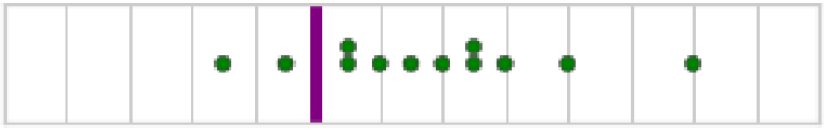


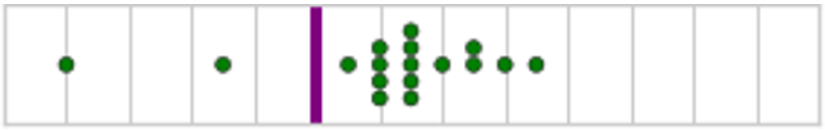
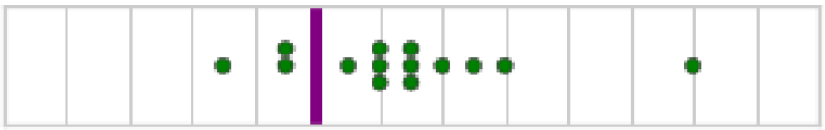
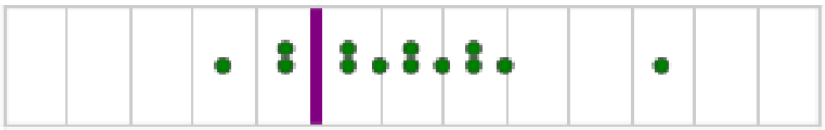
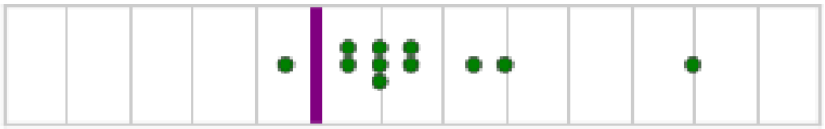
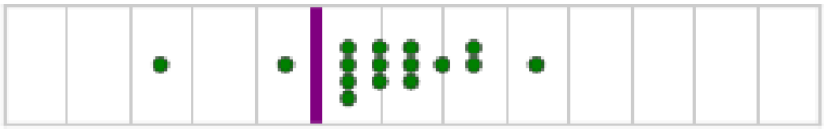
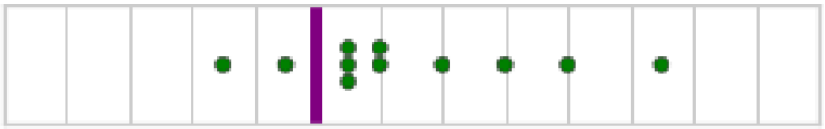
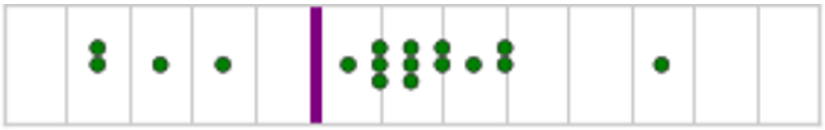
#	Post Title	Total	Vote Spread
1	Embedded Agents	209	 (One outlier vote of +17 is not shown)
2	The Rocket Alignment Problem	183	
3	Local Validity as a Key to Sanity and Civilization	133	
4	Arguments about fast takeoff	98	
5	The Costly Coordination Mechanism of Common Knowledge	95	
6	Toward a New Technical Explanation of Technical Explanation	91	

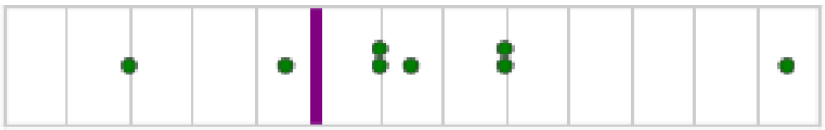
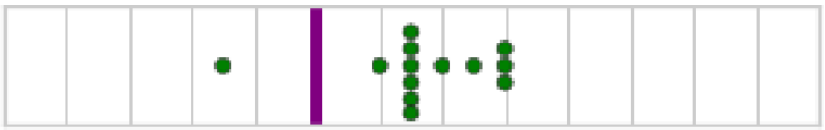
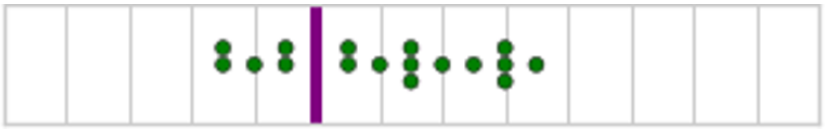
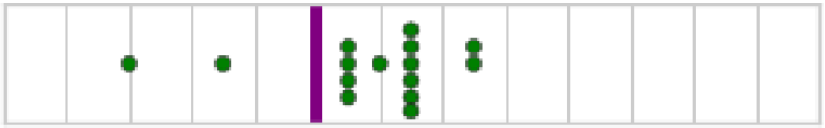
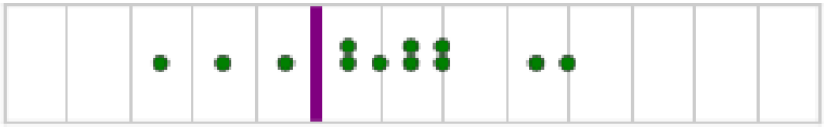
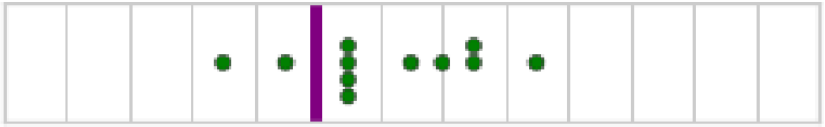
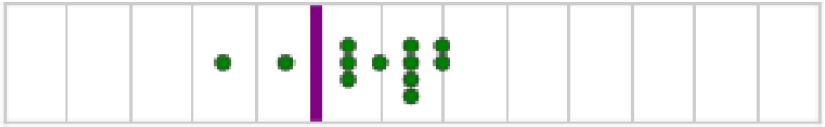
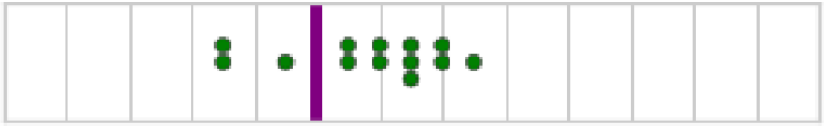
#	Post Title	Total	Vote Spread
7	Anti-social Punishment	90	 (One outlier vote of +20 is not shown)
8	The Tails Coming Apart As Metaphor For Life	89	
9	Babble	85	
10	The Loudest Alarm Is Probably False	84	
11	The Intelligent Social Web	79	
12	Prediction Markets: When Do They Work?	77	
13	Coherence arguments do not imply goal-directed behavior	76	
14	Is Science Slowing Down?	75	
15	Robustness to Scale	74	

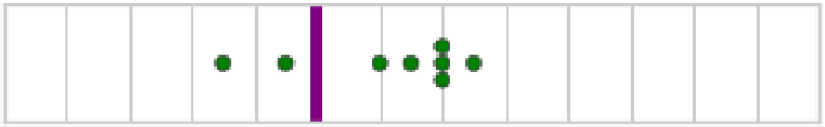
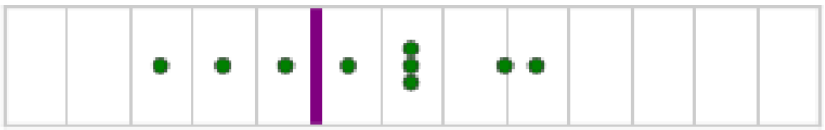
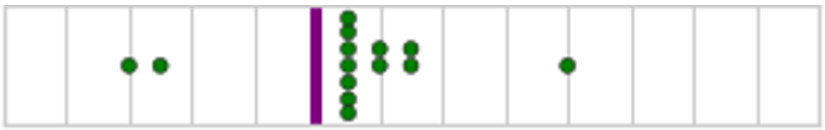
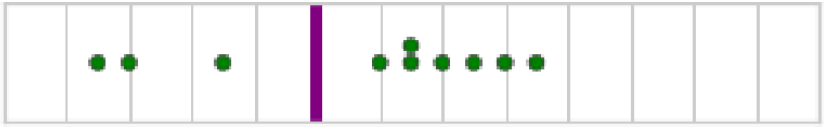
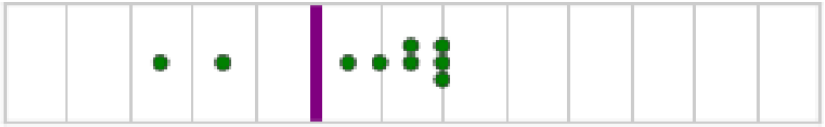
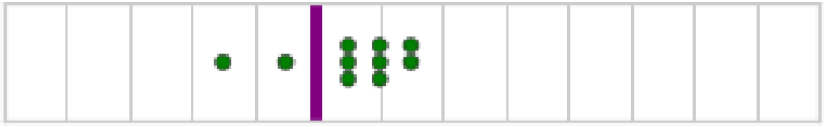
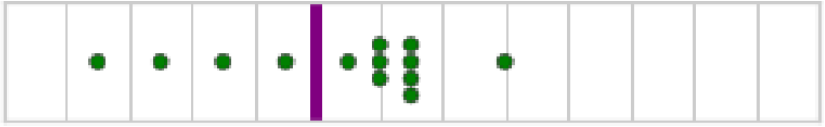
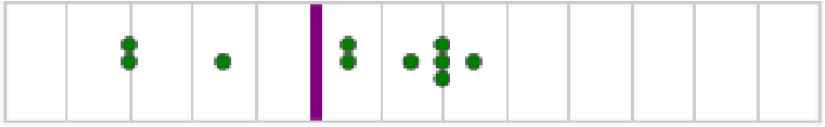
#	Post Title	Total	Vote Spread
15	A voting theory primer for rationalists	74	
17	Toolbox-thinking and Law-thinking	73	
18	A Sketch of Good Communication	72	
19	A LessWrong Crypto Autopsy	71	
20	Paul's research agenda FAQ	70	
21	Unrolling social metacognition: Three levels of meta are not enough.	69	
22	An Untrollable Mathematician Illustrated	65	
23	Specification gaming examples in AI	64	
23	Will AI See Sudden Progress?	64	

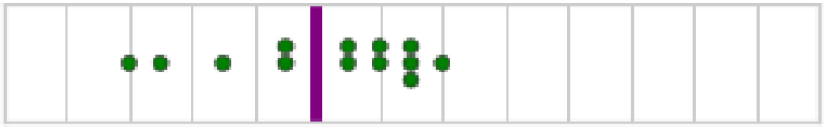
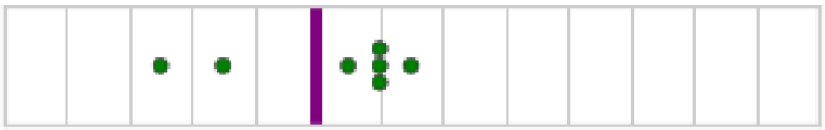
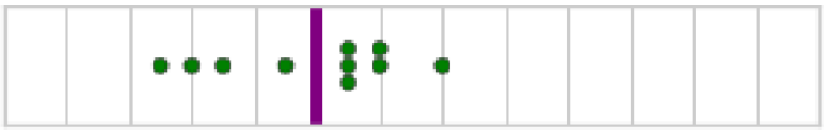
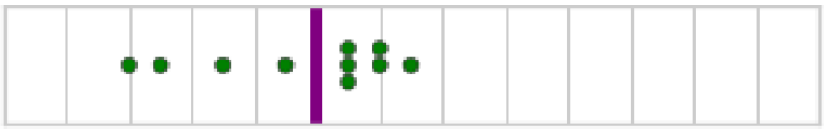
#	Post Title	Total	Vote Spread
23	Varieties Of Argumentative Experience	64	
26	Meta-Honesty: Firming Up Honesty Around Its Edge-Cases	62	
27	My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms	60	
27	Naming the Nameless	60	
27	Inadequate Equilibria vs. Governance of the Commons	60	
30	2018 AI Alignment Literature Review and Charity Comparison	57	
31	Noticing the Taste of Lotus	55	
31	On Doing the Improbable	55	
31	The Pavlov Strategy	55	

#	Post Title	Total	Vote Spread
31	Being a Robust, Coherent Agent (V2)	55	
35	Spaghetti Towers	54	
36	Beyond Astronomical Waste	51	
36	Research: Rescuers during the Holocaust	51	
38	Open question: are minimal circuits daemon-free?	48	
38	Decoupling vs Contextualising Norms	48	 (One outlier vote of +23)
40	On the Loss and Preservation of Knowledge	47	
41	Is Clickbait Destroying Our General Intelligence?	46	
42	What makes people intellectually active?	43	

#	Post Title	Total	Vote Spread
43	Why everything might have taken so long	40	
44	Challenges to Christiano's capability amplification proposal	39	
45	Public Positions and Private Guts	38	
46	Clarifying "AI Alignment"	36	
46	Expressive Vocabulary	36	
48	Bottle Caps Aren't Optimisers	34	
49	Argue Politics* With Your Best Friends	32	
50	Player vs. Character: A Two-Level Model of Ethics	30	
51	Conversational Cultures: Combat vs Nurture (V2)	29	

#	Post Title	Total	Vote Spread
51	Act of Charity	29	
53	Optimization Amplifies	27	
53	Circling	27	 (One outlier vote of -17)
55	Realism about rationality	25	 (Two outliers of -30 and +18)
55	Caring less	25	
57	Lessons from the Cold War on Information Hazards: Why Internal Communication is Critical	24	
57	The Bat and Ball Problem Revisited	24	
59	Argument, intuition, and recursion	21	
59	Unknown Knowns	21	

#	Post Title	Total	Vote Spread
61	Competitive Markets as Distributed Backprop	18	
62	Towards a New Impact Measure	14	
62	Explicit and Implicit Communication	14	
62	On the Chatham House Rule	14	
62	Historical mathematicians exhibit a birth order effect too	14	
66	Everything I ever needed to know, I learned from World of Warcraft: Goodhart's law	13	
67	The funnel of human experience	11	
68	Understanding is translation	9	
69	Preliminary thoughts on moral weight	7	

#	Post Title	Total	Vote Spread
70	Metaphilosophical competence can't be disentangled from alignment	3	
71	Two types of mathematician	2	
72	How did academia ensure papers were correct in the early 20th Century?	-2	
73	Birth order effect found in Nobel Laureates in Physics	-5	
74	Give praise	-10	
75	Affordance Widths	-142	

(One outlier of -29)

How reliable is the output of this vote?

For most posts, between 10-20 people voted on them (median of 17). A change by 10-15 in a post's score is enough to move a post up or down around 10 positions within the rankings. This is equal to a few moderate strength votes from two or three people, or an exceedingly strong vote from a single strongly-feeling voter. This means that the system is somewhat noisy, though it seems to me very unlikely that posts at the very top could end up placed much differently.

The vote was also affected by two technical mistakes the team made:

1. *The post-order was not randomized.* For the first half of the voting period, the posts on the voting page appeared in order of number of nominations (least to most) instead of appearing randomly, thereby giving more visual attention to the first ~15 or so posts (these were posts with 2 nominations). Ruby looked into it and says that 15-30% more people cast votes on these earlier-appearing posts compared to those appearing elsewhere in the list. *Thanks to gjm for identifying this issue*.
2. *Users were given some free negative votes.* When calculating the cost of users' votes, we used a simple equation, but missed that it produced an off-by-one error for negative numbers. Essentially, users got a free 1-negative-vote-weight on all the posts to which they had voted on negatively. To correct for this, for those who had exceeded their budget - 18 users in total - we reduced the strength of their negative votes by a single unit, and for those who had not spent all their points their votes were unaffected. This didn't affect the rank-ordering very much, a few posts changed by 1 position, and a smaller number changed by 2-3 positions.

The effect size of these errors is not certain since it's hard to know how people would have voted counterfactually. My sense is that the effect is pretty small, and that the majority of noise in the system comes from elsewhere.

Finally, we discarded exactly one ballot, which spent 10,000 points on voting instead of the allotted 500. Had a user gone over by a small amount e.g. 1-50 points, we had planned to scale their votes down to fit the budget. However when someone's allocation

was so extreme, we were honestly unsure what adjustment to their votes they would have wanted, as if their points had been normalised down to 500, the majority of their votes would have been adjusted to zero. (This decision was made without knowing the user who cast the ballot or which posts were affected.)

Overall, I think the vote is a good indicator to about 10 places within the rankings, but, for example, I wouldn't agonise over whether a post is at position #42 vs #43.

The Future

This has been the first LessWrong Annual Review. This project was started with the vision of creating a piece of infrastructure that would:

1. Create common knowledge about how the LessWrong community feels about various posts and topics and the progress we've made.
2. Improve our longterm incentives, feedback, and rewards for authors.
3. Help create a highly curated "Best of 2018" Sequence and Book.

The vote reveals much disagreement between LessWrongers. Every post has at least five positive votes and every post had at least one negative vote – except for An Untrollable Mathematician Illustrated by Abram Demski, which was evidently just too likeable – and many people had strongly different feelings about many posts. Many of these seem more interesting to me than the specific ranking of the given post.

In total, users wrote 207 nominations and 120 reviews, and many authors updated their posts with new thinking, or clearer explanations, showing that both readers and authors reflected a lot (and I think changed their mind a lot) during the review period. I think all of this is great, and like the idea of us having a Schelling time in the year for this sort of thinking.

Speaking for myself, this has been a fascinating and successful experiment - I've learned a lot. My thanks to Ray for pushing me and the rest of the team to actually do it this year, in a move-fast-and-break-things kind of way. The team will be conducting a *Review of the Review* where we take stock of what happened, discuss the value and costs of the Review process, and think about how to make the review process more effective and efficient in future years.

In the coming months, the LessWrong team will write further analyses of the vote data, award prizes to authors and reviewers, and use the vote to help design a sequence and a book of the best writing on LW from 2018.

I think it's awesome that we can do things like this, and I was honestly surprised by the level of community participation. Thanks to everyone who helped out in the LessWrong 2018 Review - everyone who nominated, reviewed, voted and wrote the posts.

Moral public goods

Automatically crossposted

Suppose that a kingdom contains a million peasants and a thousand nobles, and:

- Each noble makes as much as 10,000 peasants put together, such that collectively the nobles get 90% of the income.
- Each noble cares about as much about themselves as they do about all peasants put together.
- Each person's welfare is logarithmic in their income.

Then it's simultaneously the case that:

1. Nobles prefer to keep money for themselves rather than donate it to peasants—money is worth 10,000x as much to a peasant, but a noble cares 1,000,000 times less about the peasant's welfare.
2. Nobles prefer a 90% income tax that is redistributed equally—a tax that costs a particular noble $1 generates $1000 of value for peasants, since all other nobles will also pay the higher taxes. That makes it a much better deal for the nobles (until the total income of nobles is roughly equal to the total income of peasants).

In this situation, let's call redistribution a "moral public good." The nobles are altruistic enough that they prefer it if everyone gives to the peasants, but it's still not worth it for any given noble to contribute anything to the collective project.

The rest of the post is about some implications of taking moral public good seriously.

1. Justifying redistribution

This gives a very strong economic argument for state redistribution: it can easily be the case that *every individual* prefers a world with high redistribution to the world with low redistribution, rich and poor alike. I think "everyone prefers this policy" is basically the strongest argument you can make on its behalf.

(In fact some people just don't care about others and so not *everyone* will benefit. I'd personally be on board with the purely selfish people just not funding redistribution, but unfortunately you can't just ask people if they want to pay more taxes and I'm not going to sweat it that much if the most selfish people lose out a little bit.)

I think this argument supports levels of redistribution like 50% (or 30% or 70% or whatever), rather than levels of redistribution like 99% that could nearly level the playing field or ensure that no billionaires exist. I think this enough to capture the vast majority of the possible benefits from redistribution, e.g. they could get most households to >50% of the average consumption.

This argument supports both foreign aid and domestic redistribution, but the foreign aid component may require international coordination. For example, if everyone in developed countries cared equally about themselves, their country, and the world, then you might end up with optimal domestic policies allocating 10% of their redistribution abroad (much less in smaller countries who have minimal influence on

global poverty, a little bit more in the US), whereas everyone would prefer a multilateral commitment to spend 50% of their redistribution abroad.

2. There are lots of public goods

I think it makes sense for states to directly fund moral public goods like existential risk mitigation, exploration, ecological preservation, arts and sciences, animal welfare improvements, *etc*. In the past I've thought it usually made more sense to just give people money and let them decide how to spend it. (I still think states and philanthropists should more often give people cash, I just now think the presumption is less strong.)

In fact, I think that at large scales (like a nation rather than a town) moral public goods are probably the majority of public goods. Caring slightly more about public goods slightly changed my perspective on the state's role. It also makes me significantly more excited about mechanisms like quadratic funding for public goods.

I enjoyed David Friedman's *The Machinery of Freedom*, but it repeats the common libertarian line that donations can help the poor just as well as taxes:

> If almost everyone is in favor of feeding the hungry, the politician may find it in his interest to do so. But, under those circumstances, the politician is unnecessary: some kind soul will give the hungry man a meal anyway. If the great majority is against the hungry man, some kind soul among the minority still may feed him—the politician will not.

This seems totally wrong. The use of coercive force is an active ingredient in the state feeding the hungry, as it is with other public good provision. Anarchists either need to make some speculative proposal to fund public goods (the current menu isn't good!) or else need to accept the pareto inefficiency of underfunding moral public goods like redistribution.

3. Altruism is not about consequentialism

Consequentialism is a really bad model for most people's altruistic behavior , and especially their compromises between altruistic and selfish ends. To model someone as a thoroughgoing consequentialist, you have two bad options:

1. They care about themselves >10 million times as much as other people. Donating to almost anything is in insane, no way the recipient values the money 10 million times more than I do.
2. They care about themselves <1% as much as everyone else in the whole world put together. When choosing between possible worlds, they would gladly give up their whole future in order to make everyone else's life a little better. Their personal preferences are nearly irrelevant when picking policies. If they found themselves in a very powerful position they would become radically more altruistic.

I think neither of these is a great model. In fact it seems like people care a lot about themselves and those around them, but at the same time, they are willing to donate small amounts of their income.

You could try to frame this as “no one is altruistic, it’s just a sham” or “people are terrible at morality.” But I think you’ll understand most people’s altruism better if you think about it as part of a collective action or public goods provision problem. People want to e.g. see a world free from extreme poverty, and they are (sometimes) willing to chip in a small part of that vision for the same reason that they are willing to chip in to the local public park—even though the actual consequence of their donation is too small for them to care much about it.

On this perspective, donating to local charities is on much more even footing with donating to distant strangers. Both are contributions to public goods, just at different scales and of different types, and that’s the thing that most unifies the way people approach and think about them. The consequentialist analysis is still relevant—helping the poor is only a moral public good because of the consequences—but it’s not that the local charity is just a consequentialist error.

In addition to misunderstanding normal humans, I think consequentialists sometimes make related errors in their own judgments. If a bunch of utilitarians want to enjoy a nice communal space, it’s worthwhile for each of them to help fund it even though it neither makes sense on utilitarian grounds nor for their own self-interests. That’s a good norm that can leave every utilitarian better off than if they’d spent the same money selfishly. I think that a lot of moral intuition and discourse is about this kind of coordination, and if you forget about that then you will both be confused by normal moral discourse and also fail to solve some real problems that everyday morality is designed to solve.

On hiding the source of knowledge

This is a linkpost for https://unstableontology.com/2020/01/26/on-hiding-the-source-of-knowledge/

I notice that when I write for a public audience, I usually present ideas in a modernist, skeptical, academic style; whereas, the way I come up with ideas is usually in part by engaging in epistemic modalities that such a style has difficulty conceptualizing or considers illegitimate, including:

- Advanced introspection and self-therapy (including focusing and meditation)
- Mathematical and/or analogical intuition applied everywhere with only spot checks (rather than rigorous proof) used for confirmation
- Identity hacking, including virtue ethics, shadow-eating, and applied performativity theory
- Altered states of mind, including psychotic and near-psychotic experiences
- Advanced cynicism and conflict theory, including generalization from personal experience
- Political radicalism and cultural criticism
- Eastern mystical philosophy (esp. Taoism, Buddhism, Tantra)
- Literal belief in self-fulfilling prophecies, illegible spiritual phenomena, etc, sometimes with decision-theoretic and/or naturalistic interpretations

This risks hiding where the knowledge actually came from. Someone could easily be mistaken into thinking they can do what I do, intellectually, just by being a skeptical academic.

I recall a conversation I had where someone (call them A) commented that some other person (call them B) had developed some ideas, then *afterwards* found academic sources agreeing with these ideas (or at least, seeming compatible), and cited these as sources in the blog post write-ups of these ideas. Person A believed that this was importantly bad in that it hides where the actual ideas came from, and assigned credit for them to a system that did not actually produce the ideas.

On the other hand, citing academics that agree with you is helpful to someone who is relying on academic peer-review as part of their epistemology. And, similarly, offering a rigorous proof is helpful for convincing someone of a mathematical principle they aren't already intuitively convinced of (in addition to constituting an extra check of this principle).

We can distinguish, then, the source of an idea from the presented epistemic justification of it. And the justificatory chain (to a skeptic) doesn't have to depend on the source. So, there is a temptation to simply present the justificatory chain, and hide the source. (Especially if the source is somehow embarrassing or delegitimized)

But, this creates a distortion, if people assume the justificatory chains are representative of the source. Information consumers may find themselves in an environment where claims are thrown around with various justifications, but where they would have quite a lot of difficulty coming up with and checking similar claims.

And, a lot of the time, the source *is* important in the justification, because the source was the original reason for privileging the hypothesis. Many things can be partially rationally justified without such partial justification being sufficient for credence,

without also knowing something about the source. (The problems of skepticism in philosophy in part relate to this: "but you have the intuition too, don't you?" only works if the other person has the same intuition (and admits to it), and arguing without appeals to intuition is quite difficult)

In addition, even if the idea is justified, *the intuition itself* is an artifact of value; knowing abstractly that "X" does not imply the actual ability to, in real situations, quickly derive the implications of "X". And so, sharing the source of the original intuition is helpful to consumers, if it can be shared. Very general sources are even more valuable, since they allow for generation of new intuitions on the fly.

Unfortunately, many such sources can't easily be shared. Some difficulties with doing so are essential and some are accidental. The essential difficulties have to do with the fact that teaching is hard; you can't assume the student already has the mental prerequisites to learn whatever you are trying to teach, as there is significant variation between different minds. The accidental difficulties have to do with social stigma, stylistic limitations, embarrassment, politics, privacy of others, etc.

Some methods for attempting to share such intuitions may result in text that seems personal and/or poetic, and be out of place in a skeptical academic context. This is in large part because such text isn't trying to justify itself by the skeptical academic standards, and is nevertheless attempting to communicate something.

Noticing this phenomenon has led me to more appreciate forewards and prefaces of books. These sections often discuss more of the messiness of idea-development than the body of the book does. There may be a nice stylistic way of doing something similar for blog posts; perhaps, an extended bibliography that includes free-form text.

I don't have a solution to this problem at the moment. However, I present this phenomenon as a problem, in the spirit of discussing problems before proposing solutions. I hope it is possible to reduce the accidental difficulties in sharing sources of knowledge, and actually-try on the essential difficulties, in a way that greatly increases the rate of interpersonal model-transfer.

AI Alignment 2018-19 Review

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

Preamble

What this post is

This is a review post of public work in AI alignment over 2019, with some inclusions from 2018. It has this preamble (~700 words), a short version / summary (~1.6k words), and a long version (~8.3k words). It is available as a Google Doc here.

There are many areas of work that are relevant to AI alignment that I have barely touched on, such as interpretability, uncertainty estimation, adversarial examples, and assured autonomy, primarily because I have not been following these fields and wouldn't be able to write a good summary of what has happened in them. I have also mostly focused on articles that provide some conceptual insight, and excluded or briefly linked to papers that primarily make quantitative improvements on important metrics. While such papers are obviously important (ultimately, our techniques need to work *well*), there isn't much to say about them in a yearly review other than that the quantitative metric was improved.

Despite these exclusions, there was still a ton of work to select from, perhaps around ~500 articles, of which over 300 have been linked to in this post. There are many interesting articles that I really enjoyed that get only a sentence of description, in which I ignore many of the points that the article makes. Most have been summarized in the Alignment Newsletter, so if you'd like to learn more about any particular link, but don't want to read the entire thing, just search for its title in the database.

What you should know about the structure of this post

I am *not* speaking for myself; by default I am trying to explain what has been said, in a way that the authors of the articles would agree with. Any extra opinion that I add will be in italics.

As a post, this is meant to be read sequentially, but the underlying structure is a graph (nodes are posts, edges connect posts that are very related). I arranged it in a sequence that highlights the most salient-to-me connections. This means that the order in which I present subtopics is very much *not* a reflection of what I think is most important in AI safety: in my presentation order, I focused on *edges* (connections) rather than *nodes* (subtopics).

Other minor details:

1. Any links from earlier than 2018 will have their year of publication right after the link (except for articles that were reposted as part of Alignment Forum sequences).
2. I typically link to blog posts; in several cases there is also an associated paper that I have not linked.

How to read this post

I have put the most effort into making the prose of the long version read smoothly. The hierarchical organization is comparatively less coherent; this is partly because I optimized the prose, and partly because AI safety work is hard to cluster. As a result, for those willing to put in the effort, I'd recommend reading the long version directly, without paying too much attention to the hierarchy. If you have less time, or are less interested in the minutiae of AI alignment research, the short version is for you.

Since I don't name authors or organizations, you may want to take this as your opportunity to form beliefs about which arguments in AI alignment are important based on the ideas (as opposed to based on trust in the author of the post).

People who keep up with AI alignment work might want to know which posts I'm referencing as they read, which is a bit hard since I don't name the posts in the text. If this describes you, you should be reading this post on the Alignment Forum, where you can hover over most links to see what they link to. Alternatively, the references section in the Google Doc lists all links in the order that they appear in the post, along with the hierarchical organization, and so you can open the references in a new tab, and read through the post and the references together.

I expect that if you aren't already familiar with them, some articles will sound crazy from my summary here; please read at least the newsletter summary and ideally the full article before arguing that it's crazy.

Acknowledgements

Thanks to the Alignment Newsletter team, Ben Pace, Oliver Habryka, Jonathan Uesato, Tom Everitt, Luke Muehlhauser, Jan Leike, Rob Bensinger, Adam Gleave, Scott Emmons, Rachel Freedman, Andrew Critch, Victoria Krakovna, and probably a few others (I really should have kept better track of this). Thanks especially to Ben Pace for suggesting that I write this review in the first place.

Short version (~1.6k words)

While the full text tries to accurately summarize different points of view, that is not a goal in this summary. Here I simply try to give a sense of the topics involved in the discussion, without saying what discussion actually happened.

Basic analysis of AI risk. Traditional arguments for AI risk argue that since agentic AI systems will apply lots of optimization, they will lead to extreme outcomes that can't be handled with normal engineering efforts. Powerful AI systems will not have their resources stolen from them, which by various dutch book theorems implies that they must be expected utility maximizers; since expected utility maximizers are goal-directed, they are dangerous.

However, the VNM theorem does not justify the assumption that an AI system will be goal-directed: such an assumption is really based on intuitions and conceptual arguments (which are still quite strong).

Comprehensive AI Services (CAIS) challenges the assumption that we will have a single agentic AI, instead suggesting that any task will be performed by a collection of

modular services.

That being said, there are several other arguments for AI risk, such as the argument that AI might cause "lock in" which may require us to solve hard philosophical problems before the development of AGI.

Nonetheless, there are disjunctive reasons to expect that catastrophe does not occur: for example, there may not be a problem, or ML researchers may solve the problem after we get "warning shots", or we could coordinate to not build unaligned AI.

Agency and optimization. One proposed problem is that of mesa optimization, in which an optimization algorithm used to train an AI creates an agent that is *itself* performing optimization. In such a scenario, we need to ensure that the "inner" optimization is also aligned.

To better understand these and other situations, it would be useful to have a formalization of optimization. This is hard: while we don't want optimization to be about our beliefs about a system, if we try to define it mechanistically, it becomes hard to avoid defining a bottle cap as an optimizer of "water kept in the bottle".

Understanding agents is another hard task. While agents are relatively well understood under the Cartesian assumption, where the agent is separate from its environment, things become much more complex and poorly-understood when the agent is a part of its environment.

Value learning. Building an AI that learns all of human value has historically been thought to be very hard, because it requires you to decompose human behavior into the "beliefs and planning" part and the "values" part, and there's no clear way to do this.

Another way of looking at it is to say that value learning requires a model that separates the given data into that which actually achieves the true "values" and that which is just "a mistake", which seems hard to do. In addition, value learning seems quite fragile to mis-specification of this human model.

Nonetheless, there are reasons for optimism. We could try to build an *adequate* utility function, which works well enough for our purposes. We can also have *uncertainty* over the utility function, and update the belief over time based on human behavior. If everything is specified correctly (a big if), as time goes on, the agent would become more and more aligned with human values. One major benefit of this is that it is *interactive* -- it doesn't require us to specify everything perfectly ahead of time.

Robustness. We would like our agents to be robust - that is, they shouldn't fail catastrophically in situations slightly different from the ones they were designed for. Within reinforcement learning, safe reinforcement learning aims to avoid mistakes, even during training. This either requires analytical (i.e. not trial-and-error) reasoning about what a "mistake" is, which requires a formal specification of what a mistake is, or an overseer who can correct the agent before it makes a mistake.

The classic example of a failure of robustness is adversarial examples, in which a tiny change to an image can drastically affect its classification. Recent research has shown that these examples are caused (at least in part) by real statistical correlations that generalize to the test set, that are nonetheless fragile to small changes. In addition, since robustness to one kind of adversary doesn't make the classifier robust to other kinds of adversaries, there has been a lot of work done on improving adversarial

evaluation in image classification. We're also seeing some of this work in reinforcement learning.

However, asking our agents to be robust to arbitrary mistakes seems to be too much -- humans certainly don't meet this bar. For AI safety, it seems like we need to ensure that our agents are robustly intent aligned, that is, they are always "trying" to do what we want. One particular way that our agents could be intent aligned is if they are corrigible, that is, they are trying to keep us "in control". This seems like a particularly easy property to verify, as conceptually it seems to be independent of the domain in which the agent is deployed.

So, we would like to ensure that even in the worst case, our agent remains corrigible. One proposal would be to train an adversary to search for "relaxed" situations in which the agent behaves incorrigibly, and then train the agent not to do that.

Scaling to superhuman abilities. If we're building corrigible agents using adversarial training, our adversary should be more capable than the agent that it is training, so that it can find all the situations in which the agent behaves incorrigibly. This requires techniques that scale to superhuman abilities. Some techniques for this include iterated amplification and debate.

In iterated amplification, we start with an initial policy, and alternate between amplification and distillation, which increase capabilities and efficiency respectively. This can encode a range of algorithms, but often amplification is done by decomposing questions and using the agent to answer subquestions, and distillation can be done using supervised learning or reinforcement learning.

In debate, we train an agent through self-play in a zero-sum game in which the agent's goal is to "win" a question-answering debate, as evaluated by a human judge. The hope is that since each "side" of the debate can point out flaws in the other side's arguments, such a setup can use a human judge to train far more capable agents while still incentivizing them to provide honest, true information.

Both iterated amplification and debate aim to train an agent that approximates the answer that one would get from an exponentially large tree of humans deliberating. The factored cognition hypothesis is that this sort of tree of humans is able to do any task we care about. This hypothesis is controversial: many have the intuition that cognition requires large contexts and flashes of intuition that couldn't be replicated by a tree of time-limited humans.

Universality. One property we would hope to have is that if we use this tree of humans as an overseer for some simpler agent, then the tree would "know everything the agent knows". If true, this property could allow us to build a significantly stronger conceptual argument for safety. It is also very related to...

Interpretability. While interpretability can help us know what the agent knows, and what the agent would do in other situations (which can help us verify if it is corrigible), there are other uses for it as well: in general, it seems better if we can understand the things we're building.

Impact regularization. While relative reachability and attainable utility preservation were developed last year, this year saw them be unified into a single framework. In addition, there was a new proposed definition of impact: change in our ability to get what we want. This notion of impact depends on knowing the utility function U. However, we might hope that we can penalize some "objective" notion, perhaps

"power", that occurs regardless of the choice of U, for the same reasons that we expect instrumental convergence.

Causal modeling. Causal models have been used recently to model the incentives for an agent under different AI safety frameworks, and to argue that by evaluating plans with the current reward function, you can remove the incentive for an agent to tamper with its reward function.

Oracles. Even if oracles are trying to maximize predictive accuracy, they could "choose" between different self-confirming predictions. We could avoid this using counterfactual oracles, which make predictions conditioning that their predictions do not influence the future.

Decision theory. There was work on decision theory, that I haven't followed very much.

Forecasting. Several resources were developed to enable effective group forecasting, including an AI forecasting dictionary that defines terms, an AI resolution council whose future opinions can be predicted, and a dataset of well-constructed exemplar questions about AI.

Separately, the debate over takeoff speeds continued, with two posts arguing forcefully for continuous takeoff, without much response (although many researchers do not agree with them). The continuity of takeoff is relevant for but doesn't completely determine whether recursive self improvement will happen, or whether some actor acquires a decisive strategic advantage. The primary implication of the debate is whether we should expect that we will have enough time to react and fix problems as they arise.

It has also become clearer that recent progress in AI has been driven to a significant degree by increasing the amount of compute devoted to AI, which suggests a more continuous takeoff. You could take the position that current methods can't do <property X> (say, causal reasoning), and so it doesn't matter how much compute you use.

AI Progress. There was a lot of progress in AI.

Field building. There were posts aiming to build the field, but they were all fairly disjointed.

The **long version** (~8.3k words) starts here.

Basic analysis of AI risk

Agentic AI systems

Much of the foundational writing about AI risk has focused on agentic AI systems. This approach (recently discussed in the post and comments here) argues that since AI agents will be exerting a lot of optimization, there will be extreme outcomes in which our regular arguments may not work. This implies that we must adopt a security mindset (2017) to ensure alignment, and it suggests that proof-level guarantees may be more important at various stages of alignment research.

Goal-directedness

The foundational writing then goes on to point out that since powerful AI systems should not be able to be dutch booked (i.e. have their resources stolen from them), they will be well modeled (2017) as expected utility maximisers. An AI system that maximizes expected utility is very likely to be dangerous. One reason was recently formalized in MDPs in which the agent gets a *random* utility function: using formalizations of power and instrumental convergence, we find some suggestive results that agents seek control over their future (from which we might infer that they will try to wrest that control from us).

However, it is not mathematically necessary that AI systems will have utility functions (except in a vacuous sense), and while there are intuitive and conceptual reasons to think that we will build goal-directed agents by default, there are alternative pathways that might be taken instead, and that are valuable to explore and build out to ensure AI safety.

This challenge to the usual argument for utility maximizers has prompted a series of articles exploring other variants of the argument, for example by restricting the class of utility functions to make it non-vacuous, or by saying that optimization processes in general will lead to goal-directed agents.

Comprehensive AI Services

Comprehensive AI Services (CAIS) also takes issue with the model of a single AGI agent hyper-competently pursuing some goal, and instead proposes a model in which different tasks are solved by specialized, competing *AI services*. This is suggesting that modularity across tasks is sufficiently useful that it will apply to AI, in the same way that it applies to humans (e.g. I have specialized in AI research, and not plumbing). The aggregate of all the services can accomplish any task, including the development of new services, making it comprehensive (analogous to the “general” in AGI). Since AI services can also do basic AI R&D research, which leads to improvement in AI services generally, we should expect recursive *technological* improvement (as opposed to recursive *self* improvement). Note that CAIS does not necessarily suggest we will be *safe*, just that the traditional risks are not as likely as we may have thought, while other emergent risks are perhaps greater.

Critics often argue that end-to-end training and integrated agent-like architectures are likely to (eventually) outperform modular services. However, through coordination services can also be integrated. In addition, this post argues that this criticism mirrors old concerns that under capitalism firms will become too large -- a concern that the post argues did not pan out.

CAIS *does* allow for AI systems that are capable of *learning* across many domains: it simply argues that these AI systems will specialize for efficiency reasons, and so will only be *competent* at a small subset of domains. This decomposition of intelligence into learning + competence has been used to explain the variation in human abilities.

(This conversation is related to much prior conversation on Tool AI, which is listed here.)

Arguments for AI risk

There are many arguments for AI risk, with each of these posts providing a list of such arguments. It is unclear whether from an outside perspective this should be taken as evidence *against* AI risk (since different researchers believe different arguments and are aiming for different "success stories") or as evidence *for* AI risk (because there are so many different sources of AI risk).

One argument that saw a lot of discussion was that we must figure out philosophy since the creation of AGI might "lock in" philosophical ideas. For example, we might not want to have AI systems with utility functions because of impossibility results in population ethics that suggest that every utility function would lead to some counterintuitive conclusion. Similarly, there are many proposals for how to define values; it may be necessary to figure out the right definition ahead of time. Rather than solving these problems directly, we could solve metaphilosophy, or delegate to humans who deliberate, whether idealized or real.

We might also worry that AIs will economically outcompete humans, give us technologies we aren't ready for, or amplify human vulnerabilities.

Under continuous takeoff, two scenarios have been proposed for what failure looks like. First, AI differentially improves society's capability to optimize metrics that are easy to measure, rather than ones that we actually care about. Second, AI agents could accidentally be trained to seek influence, and then fail catastrophically at some point in the future once they are sufficiently capable. One critique argues that these principal-agent problems only lead to bounded losses (i.e. they aren't catastrophic), but several others disagree.

This post argues that there has been a shift in the arguments that motivate new AI risk researchers, and calls for more explanation of these arguments so that they can be properly evaluated.

Arguments against AI risk

Many views that expect the problem to be solved by default have also been written up this year.

A series of four conversations (summarized here) suggested that some engaged people expect AI to go well by default, because they are unconvinced by the traditional arguments for AI risk, find discontinuities in AI capabilities relatively unlikely, and are hopeful that there will be "warning shots" that demonstrate problems, that the existing ML community will then successfully fix.

One post lists several good outside-view heuristics that argue against AI x-risk, while another questions why value being complex and fragile must lead to high AI risk.

This talk argues that while AGI will intuitively be a big deal, it's not obvious that we can affect its impact, and so it's not obvious that longtermists should focus on it. It gives an analogy to trying to influence the impact of electricity, before electricity was commonplace, and suggests there was little impact one could have had on its safe use. It argues that accident risks in particular draw on fuzzy, intuitive concepts, haven't been engaged with much by critics, and don't sway most AI researchers.

Despite the seeming controversy in this and previous sections, it is worth noting that there is general agreement within the AI safety community on the following broader

argument for work on AI safety:

1. Superhuman agents are not *required* to treat humans well, in the same way that humans aren't required to treat gorillas well.
2. You should have a good *technical* reason to expect that superhuman agents *will* treat humans well.
3. We do not currently have such a reason.

Agency and optimization

Mesa optimization

The problem of mesa optimization was explained in significantly more detail (see also this less formal summary). In mesa optimization, we start with a *base optimizer* like gradient descent that searches for a policy that accomplishes some complex task. For sufficiently complex tasks, it seems likely that the best policy will *itself* be an optimizer. (Meta learning is explicitly trying to learn policies that are also optimizers.) However, the policy could be optimizing a different goal, called the *mesa objective*, rather than the *base objective*.

Optimizing the mesa objective must lead to good base objective behavior on the training distribution (else gradient descent would not select it), but could be arbitrarily bad when off distribution. For example, a plausible mesa objective would be to seek influence: such an agent would initially do what we want it to do (since otherwise we would shut it down), but might turn against us once it has accumulated enough power.

This decomposes the overall alignment problem into *outer alignment* (ensuring that the base objective is aligned with "what we want") and *inner alignment* (ensuring that the mesa objective is aligned with the base objective). This is somewhat analogous to different types (2017) of Goodhart's law.

The paper and subsequent analysis identify and categorize relationships between the base and mesa objectives, and explain how mesa optimizers could fail catastrophically. Of particular interest is that mesa optimizers should be fast, but could still be misaligned, suggesting that penalizing compute is not enough to solve inner alignment.

Effectively, the concern is that our AI systems will have capabilities that generalize, but objectives that don't. Since this is what drives risk, some suggest that we should talk about this phenomenon, without needing to bring in the baggage of "optimization", a term we have yet to understand well, while others argue that even if we start with this definition, it would be useful to reintroduce the notions of optimization and agency.

One advantage of the original definition is that it specifies a particular mechanism by which risk arises; this gives us a foothold into the problem that allows us to propose potential solutions and empirical investigations. Of course, this is actively counterproductive if the risk arises by some other mechanism, but we might expect optimization to be especially likely because optimization algorithms are simple, and the phenomenon of double descent suggests that neural nets have an inductive bias towards simplicity.

What are optimization and agency, anyway?

Given the central importance of optimization to inner alignment and AI safety more broadly, we'd like to be able to formalize it. However, it's not clear how to do so: while we want optimization to be about the *mechanical process* by which outcomes happen (as opposed to e.g. *our beliefs* about that process), we cannot simply say that X is an optimizer if it makes some quantity go up: by this definition, a bottle cap would be an optimizer for "keeping water in the bottle''.

It is also relevant how the system interacts with its environment, rather than just being about whether some number is going up. The type of computation matters: while older models of optimization involve an agent that can search over possible actions and simulate their results, other optimization processes must control their environment without being able to simulate the consequences of their choice.

Our use of the word "agency" might be tied to our models or specific human architectures, rather than being a general concept that could describe a mechanical property of a computation. This would be particularly worrying since it would mean that arguments for AI risk are based on our flawed models of reality, rather than an objective property about reality. However, this is extremely speculative.

Embedded agency

Discussions about AI usually assume that a notion of the "actions" that an agent can take. However, the embedded agency sequence points out that this "Cartesian boundary" does not actually exist: since any real agent is embedded in the real world, you cannot make many assumptions that are common in reinforcement learning, such as dedicated and perfectly trusted input-output channels, a perfect model of the environment, an agent architecture that is uninfluenced by the environment, etc.

This means you can never consider all of the important information, and optimize everything that could be optimized. This has led to a couple of hypotheses:

1. Real learning algorithms require modeling assumptions to solve the credit assignment problem, and so can only lead to partial agency or myopia. (See also this parable and associated thoughts.)
2. Embedded agency works via abstraction, which is the key idea allowing you to make maps that are smaller than the territory.

Value learning

Descriptive embedded agency

While the embedded agency sequence is written from the perspective of *prescribing* how ideal agents should operate, we could also aim for a theory that can *describe* real agents like humans. This involves making your theory of agency correspondingly broader: for example, moving from utility functions to markets or subagents, which are more general. The development of such a theory is more grounded in concrete real systems, and more likely to generate theoretical insight or counterexamples, making it a good research meta-strategy.

Such a theory would be useful so that we can build AI systems that can model humans and human values while avoiding embedded agency problems with humans.

The difficulty of value learning

Even if we ignore problems of embedded agency, there are obstacles to value learning. For example, there need not be a reward function over observations that leads to what we want in POMDP (though we could instead focus on instrumental reward functions defined on states).

Another key problem is that all you ever get to observe is behavior; this then needs to be decomposed into "beliefs" and "values", but there is no clear criterion (2017) that separates them (although it hasn't been proven that simplicity doesn't work, and human priors help). This suggests that ambitious value learning, in which you identify the one true utility function, is hard.

Human models

For an agent to outperform the process generating its data, it must understand the ways in which that process makes mistakes. So, to outperform humans at a task given only human demonstrations of that task, you need to detect human mistakes in the demonstrations. Modeling humans to this fidelity is an unsolved problem, though there is a little progress, and we might hope that we can make assumptions about the structure of the model.

Any such model is likely to be misspecified, and value learning algorithms are not currently robust to misspecification: in one case, the simpler but less conceptually accurate model is more robust.

You might hope that if we give up on *outperforming* humans and just *imitate* them, this would be safe. Even this is controversial, because perhaps humans themselves are unsafe, maybe imitating humans leads to mesa optimization, or possibly perfect imitation is too hard to achieve.

You might also hope that AI systems have good enough models that you can simply provide natural language instructions and the AI does what you mean.

The presence of human models in an AI system has a few unfortunate effects:

1. We can't test an AI system by seeing if it agrees with human judgment, because the AI system may be using its human model to (in the short term) optimize for agreement with human judgment
2. A bug in the code is more likely to optimize for suffering (since the human model would include the concept of suffering)
3. If humans are modeled with sufficient fidelity, these models may themselves be conscious and capable of suffering.

Learning an adequate utility function

Despite the objections that learning values is hard, it seems like humans are pretty good at learning the values of other humans, even if not perfect. Perhaps we could

replicate this, in order to learn an *adequate* utility function that leads to okay outcomes?

The main issue is that we are only good at predicting human values in *normal* situations, while powerful AI systems will likely put us in extreme situations where we will disagree much more about values. As a result, we need a theory of human values that defines what to do in these situations. One theory, associated value learning agenda, and toy model propose that we can extract partial preferences from human mental models, and synthesize them together into a full utility function, while respecting meta-preferences about preferences and the synthesis process and taking care to properly normalize utilities.

In fact, the core pieces of such an approach seem necessary for any solution to the problem. However, this research agenda depends upon solving many hard problems explicitly in a human-understandable way, which doesn't jive with the bitter lesson that ML progress primarily happens by using more compute to solve harder problems.

I don't agree that the core pieces identified in this research agenda must be solved before creating powerful AI, nor that we must have explicit solutions to the problems.

Uncertainty over the utility function

We could also make the AI uncertain about the utility function, and ensure that it has a way to learn about the utility function that is grounded in human behavior. Then, as an instrumental goal for maximizing expected reward, the AI will choose actions with high expected information gain. While this was proposed earlier (2016), the book Human Compatible (summary, podcast 1, podcast 2, interview) explores the idea in much more detail than previous writing, and it has now made its way into deep reinforcement learning as well.

Intuitively, since the AI is uncertain about the true reward, it will behave conservatively and try to learn about the true reward, thus avoiding Goodhart's law (see also fuzziness). Of course, once the AI has learned everything there is to learn, it will behave (2015?) just like a regular utility maximizer. In this setting, you would hope that the AI has become aligned with the true utility function, as long as its initial distribution over utility functions contains the truth, and the observation model by which its distribution is updated is "correct". However, it might be quite difficult to ensure that these actually hold. This also depends on the assumption that there *is* a true utility function, and that the human *knows* it, which is not the case, though this is being addressed.

One important feature of this agenda is that rather than requiring a perfect utility function to begin with, the AI can learn the utility function by interacting with the human; such a feedback mechanism can make a problem much easier. Interaction also opens up other possibilities, such as learning human norms instead of values. However, it is computationally difficult, and so more research would be needed to make it a viable solution.

Current methods for learning human preferences

There has been a lot of practical work on learning human preferences, including:

- Building a mistake model by comparing demonstrations of varying degrees of optimality (either by getting humans to rank demonstrations or by introducing noise into optimal demonstrations)
- Learning human-interpretable representations in order to advise humans
- New forms of human guidance, such as having humans provide the advantage function (instead of the reward), following natural language instructions, and learning from the (human-optimized) initial state
- Learning preferences for natural language (rather than the standard RL setup)
- Combining multiple types of human feedback
- General improvements in imitation learning and inverse reinforcement learning.

There are many recent papers that I haven't cited here, as it is a very large area of work.

Robustness

Safe reinforcement learning

We would like to ensure that our AI systems do not make mistakes during training. With preference learning, we can do this by learning human preferences over hypothetical behaviors that are not actually executed. Another option is to provide safety constraints and ensure that the AI never violates them (even during training), or at least to significantly reduce such violations.

Avoiding *all* mistakes would require us to have a formal specification of what a "mistake" is, or to have some overseer that can identify "mistakes" before execution, so that our AI could avoid the mistake even though it hasn't seen this situation before. *This seems prohibitively hard to me if we include literally all "mistakes".*

Adversarial examples

Adversarial examples are a clear demonstration of how the "cognition" of neural nets is different from our own: by making superficial changes to the input that would not matter to a human, you can completely change the output of the neural net. While I am not an expert here, and certainly have not read the huge mountains of work done over the last year, I do want to highlight a few things.

First, while we might nominally think of adversarial examples as "bugs" in our neural net, this paper shows that image classifiers are picking up *real* imperceptible features that *do* generalize to the test set. The classifiers really are maximizing predictive accuracy; the problem is that we want them to predict labels based on the features that *we* use, instead of imperceptible (but predictive) features. Adversarial training removes these fragile features, leaving only the robust features; this makes subsequent applications easier.

While the paper was controversial, I thought that its main thesis seemed to be supported even after reading these six responses.

Second, there has been a distinct shift away from the L-infinity norm ball threat model of adversarial examples. So far, it seems that robustness to one set of perturbations doesn't grant robustness to other perturbations, prompting the development of multiple perturbations, a benchmark of natural adversarial examples, and new evaluation metrics. While the L-infinity norm ball is an interesting unsolved research problem, it is in no way a realistic threat model.

Third, adversarial attacks are now being proposed as a method for evaluating how robust an agent trained by reinforcement learning is. This seems especially important since in RL there is often no train-test split, and so it is hard to tell whether an agent has "memorized" a single trajectory or actually learned a policy that works well across a variety of circumstances.

Intent alignment

Ultimately, robustness seeks to identify and eliminate all "bugs", i.e. behaviors that are inconsistent with the specification (see also this podcast). Instead of considering all the mistakes, we could seek to only prevent catastrophic mistakes, and ensure that the AI is *intent aligned*, that is, it is always *trying* to do what we want. This goal avoids many of the pitfalls around the goal of designing an AI with the right utility function.

Corrigibility

One promising way in which an AI could be intent aligned is by being corrigible: roughly, the AI is not trying to deceive us, it clarifies its uncertainty by asking us, it learns about our preferences, it shuts down if we ask it to, etc. This is a *narrower* concept than intent alignment: an AI that infers our "true" utility function and optimizes it may wrest control away from us in order to expand faster, or make us safer; such an AI would be aligned but not corrigible. There are a few benefits of using corrigibility:

1. It can be achieved with relatively low levels of intelligence (we can imagine corrigible humans)
2. It seems to have a positive feedback loop (that is, an AI that reaches some "threshold" of corrigibility would tend to become more corrigible)
3. It doesn't seem to require any domain expertise.

(A similar idea would be to build an AI system that only takes actions that the overseer has given informed consent for.)

Note that MIRI's notion of corrigibility (2015) is similar but much stricter. My guess is that MIRI wants the same intuitive corrigibility properties, but wants them to be created by a *simple* change to the *utility function*. Simplicity helps ensure that it cannot be gamed, and the utility function means that you are changing what the AI cares about, rather than trying to constrain a powerful superintelligence. For example, I'd guess that MIRI-corrigibility can depend on whether a shutdown button is pressed, but cannot depend on the *reasons* for which the shutdown button is pressed.

If you set aside the utility function requirement, then this property can be achieved using constrained optimization: the agent can optimize normally when the button is not pressed, while ensuring that it is still able to shut down if necessary, and it can optimize for shutting down if the button is pressed. If you set aside the simplicity

requirement, then you can define the desired policies and recover the correct utility function. But from now on I'm only going to talk about the notion of corrigibility I first introduced.

It has been argued that while corrigibility is simpler than "human values", it is a "non-natural" type of cognition, such that you are unlikely to be able to find corrigible intelligences with machine learning. *(I do not feel the force of this intuition; I agree much more with the earlier intuitions.)*

You might be worried that since a corrigible AI defers to us, if we were about to take a suboptimal action that we couldn't tell was suboptimal, the AI wouldn't stop us from doing so because it can't explain to us what would be bad about the world. However, at the very least, it can say "this is bad for reasons I can't fully explain".

Worst case guarantees

We still want to guarantee that there will *never* be a failure of corrigibility, which can't be done with regular ML techniques, which only give an average-case guarantee. In order to get a worst-case guarantee, we need other techniques. One proposal is to use adversarial training to find abstracted inputs on which the agent is incorrigible, where the adversary is aided by interpretability techniques that allow the adversary to understand what the agent is thinking. It would be particularly nice to find a mechanistic description of corrigibility, as that would make it easier to verify the absence of incorrigible behavior.

Critics argue that this could never work because machine learning wouldn't learn the "intended" interpretation of corrigibility, and could be adversarial. *I don't think this objection is critical. It seems like it is saying that ML will fail to generalize and there will be situations in which the concept of corrigibility breaks down, but the entire point of adversarial training is to find these situations and train the agent away from it.*

While this is usually tied in to the broader iterated amplification agenda, it seems to me that solving just this subproblem would achieve a lot of the value of the full agenda. If we had a way of applying adversarial training to an arbitrary AI agent, such that we are very likely to find potential inputs on which the agent is incorrigible, then presumably AI systems that could be incorrigible would not be deployed. Iterated amplification adds additional safety in that it (hopefully) allows you to assume a smarter, already-aligned adversary, whereas a direct solution to this subproblem would have an approximately-as-capable, not-automatically-aligned adversary, which would probably not have a worst-case guarantee but might still be good enough.

Scaling to superhuman abilities

Iterated amplification

Iterated amplification carves out a broad class of algorithms that can scale to superhuman abilities, with the hope that we can analyze the alignment properties of the entire class of algorithms at once. Algorithms in this class have two components:

1. Amplification, which increases an agent's capabilities, at the cost of efficiency.
2. Distillation, which increases an agent's efficiency, at the cost of capability.

Given this, starting from some base agent, the algorithm alternates amplification and distillation, to get successively more capable agents, as long as each component is good enough.

Given this broad class of algorithms, we can instantiate many specific algorithms by picking a specific amplification step and a specific distillation step. For example, the amplification step can be done by allowing an overseer to *decompose* the problem into subproblems, which is especially promising for question answering. Distillation could be done using supervised learning, imitation learning, or reinforcement learning.

Recursive reward modeling (podcast) is another algorithm that could allow us to scale to superhuman abilities. It can be cast as an algorithm in the iterated amplification class by considering an amplification step that takes agents that can evaluate some set of tasks, and builds new human-agent teams that can evaluate some more complex set of tasks. The distillation step would then be reinforcement learning, to get an agent that can directly solve the more complex tasks. Iterating this eventually leads to an agent that can solve the original desired task.

Iterated amplification does impose a particular structure on algorithms, which can be applied to existing ML problems. However, this may be uncompetitive if the best ML algorithms require different algorithmic structures or different environments, in order to reach high capabilities (though we could then train a question-answering system alongside the other algorithm / environment, which plausibly doesn't take too many more resources).

The iterated amplification sequence, recursive reward modeling paper, and these posts help explain the full agenda better.

Quantilization

Quantilization (2015) allows you to amplify a base policy by randomly selecting among the top 1/Q of actions the base policy could take, at a cost of at most Q-fold increase in risk. However, this can forgo benefits of the rest of the base policy. Since quantilization increases risk, it cannot be safely iterated: for example, if you start with a policy with a worst-case 1% chance of failure, and you 5-quantilize it, you now have a worst-case 5% chance of failure. After two more iterations of 5-quantilization, there is no longer a worst-case bound on failure probability.

Debate

Another mechanism for scaling beyond humans is debate (podcast), in which an AI agent is trained via self-play in a zero-sum game in which its goal is to "win" the debate, as evaluated by a human judge. The key hope is that detecting a lie is easier than lying: if one of the players lies or deceives or manipulates the human, then the other player can reveal that and thereby win the debate. If this were true, we would expect that the equilibrium behavior is for the agent to provide honest, useful information.

Since its proposal, debate has been tested with MNIST and Fashion MNIST, as well as question answering. There is also a proposal to use it to improve iterated amplification.

Theoretical work brings up the possibility of questions that are "too hard": while sufficiently long "feature debates" are provably truth-seeking (because the debaters can reveal all of their information), it is possible to construct complex questions in which the debate doesn't find the right answer. However, the results don't generalize well from feature debates to real debates.

Relatedly, even if it is easy to detect lies, it's not clear what would happen with ambiguous questions.

Since debate doesn't involve alternating between increasing capabilities and increasing efficiency, it isn't an instance of iterated amplification. However, both iterated amplification and debate are aiming to compute the answer that an exponentially large tree of bounded humans would arrive at (see next section), and so it seems likely that either they would both work, or neither would work.

Factored cognition

Both iterated amplification and debate depend on the factored cognition hypothesis: that arbitrarily complex tasks can be performed arbitrarily well by a giant tree of bounded base agents, possibly extended with features like shared external memory or long-lived assistants (2016).

Iterated amplification checks local nodes in a tree of considerations and broken-down questions, in which an assistant at level k decomposes its questions, gets answers from assistants at level k-1, and combines them into an overall answer. Meanwhile, in debate, if the two agents disagree, they will play down the most difficult / contested path in an exponential tree of arguments and counterarguments, so the debate training procedure is checking a single path from root to leaf in the exponential tree.

It is an open question whether the factored cognition hypothesis is true. Empirical work has been scaling up, and we should hopefully have some informative evidence in the upcoming year.

The main reasons people are skeptical of the hypothesis are because it seems that sufficiently complex tasks require building up big contexts or using globally-constructed intuitions or "inexplicable flashes of insight". This could be done if the "small" agents simulated an arbitrary Turing Machine, but this would lose any guarantees of alignment.

However, we might expect that these tasks could still be done by a tree of humans: humans are allowed to use a heuristic "just because it works"; this should allow the tree of humans to use heuristics that other agents use, including "inexplicable flashes of insight".

Universality

Alignment of the tree of humans

In order for this tree of humans to be aligned (a necessary condition for iterated amplification or debate to be aligned), the initial agent must already be aligned, and putting the agents together must not destroy alignment. One intuition that this is hard

is that alignment is not compositional; a “big” agent made up of “small” aligned agents need not be aligned. However, the hope doesn’t depend on compositionality of alignment; it instead depends on ensuring that your agents never do incorrigible optimization.

In addition, it could be the case that “large” initial agents like humans (or human imitations) are not robustly aligned, because there may be some clever argument that causes them to behave incorrigibly. One response would be to use low-bandwidth overseers as the initial agent, who only answer very “small” questions on which we are relatively confident that there are no such failures. We would also hope to train humans to properly decompose questions and behave corrigibly, so that putting together several humans remains corrigible (a task for which we need social scientists).

Note that it is only competitive to approximate the tree of humans with iterated amplification if we expect that any powerful AI systems will also be trained in a manner similar to iterated amplification. If we instead consider a model in which ML perfectly optimizes a function (rather than performing iterated local search), then iterated amplification would be far more expensive than unaligned powerful AI systems. It would be worth studying this simpler model to see if alignment is possible there.

Ascription universality

Even if we know that the tree of humans is aligned, we also need to ensure that the model trained from oversight from the tree of humans will also be aligned. The key claim in favor of this is that HCH (the tree of humans) is *universal*, that is, it “knows” any facts that a sufficiently smaller computation “knows”. This was formalized here and applied to multiple problems, including the problem that malign optimization might emerge *within* HCH. While a good explanation of this is out of scope here, I summarized these posts here. Ascription universality does have to be applied to the entire training process and not just the final model.

Interpretability

Since we want to be able to “know everything the model knows”, and also to be able to find situations under with a model behaves corrigibly (see worst case guarantees above), it would be very useful to be able to peer inside our models and understand what they are doing. It would be particularly useful to be able to identify optimization processes and understand how they come about.

Even though interpretability tools probably could not deal with already deceptive models, since the deceptive models could figure out how to fool the tools, it seems likely that interpretability could help prevent deception from ever arising -- hopefully an easier task.

However, interpretability has other uses besides catching problems: it could also be used to get more understandable models during training, provide feedback on the *process* by which a model makes a decision (rather than feedback on just the decision), or create ML techniques that help us understand the world without acting in it (thus avoiding problems with agential AI).

Unfortunately, I haven't kept up with interpretability research, so I can't say how it's progressed recently, but one paper you could start with is activation atlases.

Impact regularization

Impact measures

In 2018, there was a lot of progress on proposing specific impact measures, including relative reachability and attainable utility preservation (followup, paper). These were recently unified as using similar underlying algorithms but with different "deviation measures": the former considers the change in number of reachable states, whereas the latter considers the change in attainable utility (for some set of utility functions).

These two posts summarize the work on impact (going back till 2012).

What is impact, anyway?

The Reframing Impact sequence aims to build intuitions about what we mean by "impact", and concludes that an action is impactful if it changes our ability to get what we want. Of course, this definition depends on "what we want", whereas usually with impact regularization we want something that is easy to specify. However, we might hope that impact is relatively goal-agnostic, because for most goals you need to pursue the same convergent instrumental subgoals. In particular, we might hope for a formalizable notion of *power*, that attainable utility preservation could penalize.

To better distinguish between different definitions and techniques for measuring impact, this post proposes several test cases for impact regularization.

Utility of impact measures

The mainline use case for impact regularization is to be an "additional layer of defense": if for some reason we fail to align an AI system, then hopefully there still won't be catastrophic consequences, because the AI system only takes low-impact actions. However, this may fail to work for a variety of reasons. Still, work on impact measures could be useful for deconfusion, testing protocols, temporary alignment measures, or value-neutrality verification.

Causal modeling

Causal influence diagrams help us understand what a training *process* does. Given a causal influence diagram, we can determine *observation incentives* (what an agent would like to know) and *intervention incentives* (what an agent would like to change). We can produce such diagrams for AGI safety frameworks, and analyze solutions to reward function tampering, user feedback tampering, and observation tampering. For example, it allows us to show that if the agent's plans are evaluated by the current reward, then there is no incentive for the agent to tamper with its reward function.

The variables of the diagrams represent important components of the agent and the environment (such as reward functions and dynamics models in the agent, and the user's preferences and the state of the world in the environment). Different ways of combining these into agent setups lead to different causal influence diagrams. The incentive analysis enables the designer to choose agent setups with good incentive properties.

However, the causal models themselves are not uniquely determined. For example, what counts as wireheading is relative to the stance taken towards the system and its desired goals. For example, if you define it as taking control of some "narrow measurement channel", then what is a measurement channel and what the goal is depends on modeling assumptions.

Oracles

Oracles also benefit from reasoning about causality and influences. A system that maximizes predictive accuracy ends up choosing self-confirming predictions, which can be arbitrarily bad. (This affects self-supervised learning in addition to oracles.) You might hope to avoid this by preventing the AI system from being aware of itself, but this doesn't work.

Instead, we could ensure that the oracle makes predictions conditional on the predictions not influencing anything (using randomization to do so). There are still other problems besides self-confirming predictions, such as acausal trade.

Decision theory

There's been a lot of work exploring the intuitions behind decision theory. Since I don't follow decision theory closely, I'm not going to try and summarize the conversation, and instead you get a list of posts: pro CDT, anti CDT, anti FDT, actually it all depends on counterfactuals, anti UDT because of commitment races, UDT doesn't work with AIXI, strange reasoning in Troll Bridge, a comparison across decision theories, counterfactual induction posts. There's also been some discussion of why people care about decision theory: it is useful for improving rationality, finding problems, and deconfusion.

Relatedly, this paper characterizes the decision theories of existing agents, and this post explains how "Pavlov" strategies (similar to reinforcement learning) can work well with game theory.

As we get to the end of the technical alignment section, I want to mention BoMAI, which didn't fit in any of the sections. BoMAI is an AIXI-like system that does not seek power, because it only cares about reward until the end of the episode (myopia), and during the episode it is confined to a box from which information cannot leave. Such an AI system can still be useful because there is also a human in the box, who can transmit information to the outside world after the episode has ended.

Strategy and coordination

So far I've been talking about the technical work on the alignment problem. Let's now switch to more "meta" work that tries to predict the future in order to prioritize across research topics.

Continuous vs discontinuous takeoff

A central disagreement among AI researchers is about how "quickly" AI improves once it reaches human level. Recently, the question has been distilled to whether there will be a *discontinuity* in AI capabilities. As a result, I will ask whether takeoff will be *continuous* or *discontinuous* (as opposed to *slow* or *fast*).

One operationalization of this question is whether there will be a 4-year doubling of GDP that ends before the first 1-year doubling of GDP starts. Note that continuous takeoff need not be slow: to get to 4-year doubling, you need superexponential growth. Under exponential growth, the doubling time stays fixed at its current value of a few decades. Extrapolating historical growth trends (which "supports the possibility of radical increases in growth rate") would still (probably) be compatible with this operationalization.

Two posts argue for continuous takeoff; the main argument is that continuity is very likely for properties that people care about, since lots of people are trying to make progress on the property, and it is less likely that we quickly invest much more effort into making progress on the property. So far, there has not been a compelling response, but this does not mean that researchers agree.

There has been some discussion of particular properties that make discontinuous takeoff seem more likely (though I would guess that they are not the arguments that MIRI researchers would make). For example, perhaps we just need to find the one correct architecture, which will then cause a discontinuity, but note that birds and primates have independently evolved neural architectures that both work well.

Alternatively, AI systems with different explicit utility functions could cooperate by merging to pursue a joint utility function, making them much more effective at coordination than humans, allowing them to avoid principal-agent problems that plague human corporations. This could lead to a discontinuous jump. AI systems could also build monopolies through such coordination to obtain a decisive strategic advantage.

We could also expect that just as the invention of culture and social learning by evolution allowed humans to become the dominant species very quickly (relatively speaking), similarly once AI systems are capable of social learning they may also "take off" discontinuously. However, the same argument could be taken as evidence *against* a discontinuity, since current natural language systems like GPT-2 could already be thought of as processing culture or doing social learning.

It is worth noting that questions about recursive self improvement and decisive strategic advantage do not map cleanly onto the question of takeoff speeds, though they are related. The primary reason takeoff speed is important is that it determines whether or not we will be able to respond to problems as they come up. For this purpose, it's probably better to define takeoff speed with respect to the amount of work that can be done as AI takes off, which might differ significantly from calendar time.

The importance of compute

There is a strong case that the most effective methods (so far) are the ones that can leverage more computation, and the AI-GA approach to general intelligence is predicated on this view (for example, by learning good learning environments). In fact, since the rise of deep learning in 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-*month* doubling time. It's important to note the caveat that we cannot simply increase compute: we also need good data, which is sparse in rare, unsafe situations (consider driving when a pedestrian suddenly jumps on the road). This may require human knowledge and explicit models.

Since it seems more likely that compute grows continuously (relative to a "deep insights" model), this would argue for a more continuous takeoff. However, you may expect that we still need deep insights, potentially because you think that current techniques could never lead to AGI, due to their lack of some property crucial to general intelligence (such as causal reasoning). However, for any such property, it seems that *some* neural net could encode that property, and the relevant question is how big the neural net has to be and how long it takes for local search to find the right computation.

Sociological evidence

It has recently become more common to critique the field of AI as a whole, which should (arguably) cause you to lengthen your timelines. For example, hypothesizing after the results are known makes for bad science that doesn't generalize, and research that is "reproducible" in the sense that the code can be rerun to get the same results need not have external validity. There is also a tendency for researchers to throw trial and error at problems, which means that with repeated trials by chance we can get results that look significant. It also means that researchers don't understand the systems they build; reorienting the field to focus on understanding could make our design decisions more deliberate and make it more likely that we build aligned AIs.

We should also expect that at least industry research is biased towards short timelines, since any companies that didn't argue for short timelines would be much less likely to get funding.

Meta work on forecasting

While forecasting the future is notoriously hard, collaborative and checkable forecasting is even harder. It would be nice to at least reduce the difficulty back down to "regular" forecasting. Three steps have been taken towards this:

1. People need to agree on the meaning of the terms used; an AI forecasting dictionary has been developed for this purpose.
2. In order to be checkable, questions need to be operationalized; but then it is often the case that the primary determinant of the answer to a question depends on some "distractor" feature. For example, whether we have a superhuman AI at <game> by 2025 depends a lot on who tries to make such an AI, rather than whether we have the technical ability to make such an AI. A

partial solution was to create a resolution council, and instead have questions ask about the future opinion of the resolution council.
3. This post provides advice on how to write good forecasting questions, with a database of examples.

Of course, there is still the hard problem of actually figuring out what happens in the future (and it's even hard to tell *whether* long-run forecasting is feasible). The Good Judgment Project studied practices that help with this problem, summarized here.

Another issue arises when asking members of a group (e.g. AI researchers) about outcomes that depend on actions within that group: due to the bystander effect, everyone may predict that the group will solve a problem, even though they themselves are not trying to solve the problem. So, we should instead ask people to make predictions about the proportion of members that try to solve a problem, and compare that to the proportion of members who say that they are trying to solve the problem.

AI Progress

A full update on AI progress in 2019 would be far too long, so here I'll just mention some results I found interesting, which biases towards 1. results involving "throwing compute at the problem", and 2. understanding deep learning.

Reinforcement learning

1. AlphaStar (update, discussion) become extremely good at Starcraft.
2. OpenAI Five beat the world champions at Dota, and could play cooperatively alongside humans.
3. OpenAI trained a robot to manipulate a Rubik's cube so that it could sometimes solve a jumbled cube when given the steps of the solution. See also this discussion.
4. MuZero is an evolution of AlphaZero where MCTS is applied on a *learned* world model optimized for planning, allowing it to master Atari in addition to AlphaZero's Go, Chess, and Shogi. See also this paper on instrumentally learned world models.
5. Pluribus was shown to be superhuman at *multiplayer* poker. (Note that to my knowledge it did not use deep learning, and it did not require much compute.)
6. With a complex enough hide-and-seek environment, self-play can learn qualitatively interesting behaviors.

Deep learning

1. While GPT-2 is the most well-known, there have been several large language models that are eerily good at capturing language, such as Transformer-XL and XLNet.
2. SATNet proposed a differentiable layer for neural networks that provides a strong inductive bias towards "logical reasoning", though even regular machine translation techniques work well for function integration and differential equation solving.
3. The lottery ticket hypothesis from 2018 was tested much more.

4. The double descent phenomenon was empirically validated.

Field building

While there have been a lot of field building efforts, they are relatively disjoint and not part of a conversation, and so I've summarized them in lists.

Summaries and reviews

1. This talk and multipart podcast provides an overview of approaches to technical AI alignment.
2. This post decomposes the beneficial AI problem into a tree of different subproblems (with a particular focus on the alignment problem).
3. There is of course the annual literature review and charity comparison.
4. This post identifies important hypotheses that researchers disagree about.

Agendas and prioritization

1. This doc provides an overview of the technical problems that need to be solved to align AI systems (as opposed to e.g. MIRI's deconfusion approach).
2. These posts list questions that could be tackled by philosophers and non-AI researchers respectively.
3. It would be better to bridge near- and long-term concerns about AI, to prevent the fields from "fighting" each other.
4. For s-risks, rather than looking at particular scenarios, we could focus on risk factors: properties we can intervene on to make risks less probable or less severe.

Events and news updates

1. Several conferences and workshops in 2019, including Beneficial AGI, SafeML at ICLR, AI Safety at IJCAI, and Uncertainty and Robustness at ICML.
2. There was a human-aligned AI summer school and an AI safety camp.
3. OpenAI switched to a limited-profit structure and received a $1B investment from Microsoft, while still expressing support for their charter.

The Center for Security and Emerging Technology (CSET) was founded.

References

See the Google Doc for a list of all the names and links in the text above.

Becoming Unusually Truth-Oriented

This is a post on "the basics" -- the simplest moment-to-moment attitudes one can take to orient toward truth, without any special calculations such as Fermi estimates or remembering priors to avoid base-rate neglect. At the same time, it's something almost everyone can fruitfully work on (I suspect), including myself.

Somewhat similar to track-back meditation.

Memory

Tip of the Tongue

The central claim here is that there's a special art associated with what you do when something is "on the tip of your tongue" and you can't quite remember it. Most people have the skill to some extent, but, it can be sharpened to a fine point.

Improved memory helps you become truth-oriented in a fact-oriented, detail-oriented sense. It works against inaccuracy. It also works against misspeaking, and thus propagating falsehoods.

Remembering Dreams

I first explicitly noticed the effectiveness of this technique for remembering dreams. When I wake up, I often have only one significant memory from my dreams. However, when I focus on the memory, *explicitly* naming each detail I can recall, and *gently* waiting for more, I can often unfold the memory into far, far more than I initially thought I could remember.

- Each detail you recall can open up more details.
- There's something special about explicitly naming details. I might have a general sense that there was a portal in the sky that looked a certain way, but explicitly confirming in my head that it looked as if the sky were broken glass, but at the same time the portal was perfectly round, might bring back more memories.
    - Writing things down on paper is probably a good way of making sure you're explicitly confirming each detail, if you want to go that far.
- It's also very important to sit with memories and give them time to bring something more. Sometimes there will be a rush of memories, with each new item bringing more and more. Other times, you'll be stuck. It's easy to fail at that step, assuming that no more is coming. In my experience, if you sit with the memories, avoid getting distracted, and gently ask for more, more will often come to you fairly soon. You'll surprise yourself with what you can remember.

Sometimes I don't even remember any images from the dream at all, but have a vague sense of the dream (excitement, peace, more complicated emotions). I can still sometimes recall much more if I explicitly describe the left-over feeling to myself in as much detail as possible, and sit with it patiently waiting for more.

Think of it as forming a better relationship with your memory. It's easier to wait patiently when you've had several experiences where it's paid off. Explicitly processing details of what you've remembered lets your memory know you're interested, helping to keep it engaged in searching for more (and, potentially, training it to retain more).

Eventually, if you're better calibrated, you won't have to wait 5 minutes trying fruitlessly if you really don't think you will remember. But in order to be well-calibrated about that, you have to try it sometimes.

You might be worried about confabulation. I'll talk more about that later.

Remembering Events

My claim is that this technique generalizes to any memory. Dreams might be a good practice case, especially if you don't have too many cognitively demanding distractions in the morning.

But you can try the same thing with anything. Someone I knew with especially good memory told me that he thought this was most of his skill; he might have started out with slightly above-average memory, at some point he started taking pride in his reputation for good memory. This prompted him to put effort into it, rehearsing memories much more than he otherwise would. People would then remark on his good memory, further reinforcing the behavior.

Conversations, and interactions with people generally, might make a good practice case. Many people already re-visit conversations mentally over and over (perhaps thinking of things they wish they'd said). You can treat these the same way as dreams, trying to recall as much detail as you can each time you think of them.

Of course, rehearsing certain memories again and again might not be a good thing. Watch whether you're worsening any mental problems such as depression. It may be good to couple this practice with staring into regrets and other emotionally balancing techniques, so that rehearsing memories is useful rather than intensifying emotional damage from those memories.

False Memories

Some studies about memory may give you pause.

- First of all, there is evidence that people fabricate false memories. So, how can we trust recall? Maybe trying harder to recall something actually generates false memories.
- Second, there's been some research suggesting that in some sense we "retrieve" memories (take them out of storage), and then "put them back"; and if the process is disrupted before we "put them back", we can be made to forget the memory. This suggests that memories might be altered every time they get touched, which would mean they'd last longer if we didn't think about them.

Unfortunately, forgetting is also a thing, so making memories last longer by avoiding them doesn't seem to be an option. Rehearsal is necessary for sharper memory.

Still, false memories seem like a significant concern. Memories just *seem real*. If false memories are really common and easy to create, what are we supposed to do about that?

I think the situation isn't really hopeless. I think most false memories are more like mistaken inferences. I might be sure I put my keys in my pants pocket, where I always put them. But then I might eventually recall that I put them somewhere else yesterday. What seemed like a memory was actually an inference.

As long as you're aware of these issues, I would expect that gently tugging on memories to recall more details would improve things rather than lead to more confabulation.

In my opinion, the critical turning point should be: if you are good enough at working with your memories that you have started to see through some of your own false memories, then you can start to become more confident in your judgements of which memories are real or false.

I could be wrong, of course. This is a critical question in how good/important the overall practice is.

Gendlin's Focusing

There's an obvious similarity between what I'm describing and Gendlin's Focusing. I similarly gently interact with a "felt sense" and try to name it, and iterate the process to get more detail. However, the "felt sense" is not especially located in my body the way it's described in Gendlin's focusing. It's possible that body sensations are actually involved at a subconscious level.

In any case, you may find the "gentle tugging" kind of stance useful for untangling emotions, not just recalling memories. Also, learning Focusing might help with memory and the other things I'm describing in this post?

The connection to Focusing also supports the idea that you can tell between true memories and confabulations by checking the degree of "fit" -- you have a felt sense, then you describe the thing explicitly (which gives you a better "handle" for it), then you ask yourself whether the name "resonates" with the felt sense. This is like a reality check (a theme I'll return to in the inner-sim section). But of course this isn't particularly reassuring unless you already believe that Focusing is uncovering (as opposed to confabulating) information.

Remembering Ideas

I tend to place a high value on remembering ideas. A forgotten idea is like a little death. I generally prefer the conversation norm of pausing if someone has forgotten an idea, possibly for a significant amount of time, so they can try and recover it. Ideas are important.

This habit gave me a lot of practice with tip-of-the-tongue type recollection and the "gentle tugging" technique. Practicing this stuff seems quite important for being able to do it when you need it. So I think giving yourself significant time to try and remember forgotten ideas is quite valuable if only as practice.

I think a similar sort of mental motion is involved in *developing* ideas, as well. Let's move on from the memory section...

Truth-Oriented Thinking

Developing Ideas

When you have an idea, you start with a kind of "pointer" -- a felt sense which says that there should be a think in a particular direction. You can unpack the pointer by explicitly naming things about it, checking for "fit" with the felt sense. The more you name, the easier it is to pull more details out.

Sometimes it turns out that the idea really doesn't make any sense at all; the things with the best "fit" don't actually do anything good when you explicitly spell them out. Then the felt sense changes.

To me, it feels like the felt sense traces out natural "pathways" across a "landscape" which you're exploring. An idea might be a pointer which leads to a dead end, but there's still "really a path there" -- you had it, which must mean that it was a natural thought to have in *some* sense. I take interest not just in what's true, but what the natural development of certain ideas is. This kind of attitude helps you explore alternative pathways.

Gendlin describes his notion of Focusing as involved in scientific research. It's not just about emotions. I think I'm describing the same thing here.

Inner Sim

CFAR teaches a class on "inner sim", the intuitive expectations you have. When you try to balance one object on top of another, you have an intuition about whether it will fall. If someone tells you something, you might have an intuition about whether they're lying. You can't necessarily unpack these intuitions very well. Nor are they perfectly accurate. But they are quite useful.

The surprising thing is that it seems many people don't naturally make use of their inner sims as much as they could. Let's say you're at work, and you come up with a plan for completing a project within a week. The words "planning fallacy" might come to mind, but let's set that aside and ask a different question -- does your inner sim really expect the project to be done in a week? This kind of question can give useful information surprisingly often. And if your inner sim doesn't think the plan will work, you can try and ask yourself questions like why it will fail.

So, once you've developed an idea via the methodology in the previous section, another thing you can do is ask your inner sim about the idea. Is it true? Is it real? Can it work? What do you *actually* expect?

Using gentle tugging for idea development is just as good for creating fact or fiction, so you have to add this kind of reality check.

Also, communicating with the inner sim can be a lot like communicating with memory. You can gently sit with the question "what do I actually expect?" and see what comes

up. And you similarly want to try and explicitly name what comes up; each detail of your expectations which you explicitly name can help pull more out.

Motivated Cognition

Just like we worried about false memories, we might worry about motivated cognition. Does asking your inner sim *really* provide a truth check? Does following your felt sense create a bias in what ideas you develop?

In my experience, if I'm caught up in motivated cognition, it is *literally harder to remember things which go against what I'm saying* -- it seems like I just don't remember them. But the same memory techniques which I've mentioned do help. I might not want to *say* the contrary facts once I recall them, but I can at least consciously decide that.

Similarly, I think the inner-sim checks are indeed useful in combating motivated cognition. Is it true? Is it real? What do I actually expect? What do I actually think? Giving yourself a little pause to sit with these questions can make you change your mind during an argument in a number of seconds (in my experience).

Correcting Yourself

For any of this to work during a conversation, I think you have to up your willingness to correct yourself. The thing is, if you notice *during* a conversation that you gave a false account of events, then there's going to be some consistency bias making you favor the version you've already said, and maybe some cognitive dissonance around not thinking of yourself as someone who gives false accounts.

It's not too uncommon for me to describe something with a nice "narrative logic" to it, and then remember some facts which don't fit the narrative. These additional facts may not even improve the other person's understanding of the situation -- the narrative is optimized to explain things in an understandable way, whereas the corrected detail isn't. But nowadays I try to mention them "for my own sanity" even if it doesn't make the conversation better.

If I don't do this, memory checks and reality checks will often feel counterproductive in conversations. If I'm unwilling to abandon my narrative *verbally*, then correcting it *internally* is a wasted motion which just generates a narrative/fact divide which I then have to track.

Explaining Things to Others

Just as explicitly naming things within your own head can help you pull detail out, once you *think* you understand something, explaining it to someone else can help pull a whole lot more detail out. This is probably true for memory, too.

It's not even necessarily about the interaction with the other person. Just trying to write something for someone else (and then never sharing it) can be similarly useful, whether it's a specific audience or a broad one. The need to bridge the inferential gap makes many more details feel *relevant*, which didn't feel relevant when you were explaining it to yourself.

Naturally, communicating an idea to another person is also great for uncovering problems.

This goes back to the reason why the overall technique I'm discussing works at all. Explicitly naming details of a memory helps to unpack it because what you know you know is different than what you know. You have a kind of mental illusion that you're remembering a whole conversation, but you're not really fitting all those details in short-term memory, which means you're not successfully pulling on all the associations. Similarly, you might *think* you understand something, but be unable to really explain all the details.

Gears Thinking

Gears-level thinking is like unpacking an idea with exceptionally high standards about whether you really understand it. I mentioned that explaining things to others is helpful because you "pull on" details which you wouldn't ordinarily pull on, since you think you understand them. Gears thinking doesn't literally pull on "everything", but it pulls on a lot more.

I'm afraid that someone will read that and kind of nod along without getting it. I'm not talking about just generally having higher standards. I'm talking about the moment-to-moment experience of thinking. I'm saying there's a mental stance you can take where you "stop being lazy about your thinking" -- you don't re-check really solid things like 1+1=2, but you aren't satisfied with a thought until you've really gotten *all* the details in a significant sense.

The question you ask isn't *whether* something is true; the question you ask is *exactly why* it's true. No matter how confident you are that, say, a theorem you're using holds, *you want the proof*. You're trying to see all the pieces and how they fit together.

It's like pulling out a moth-eaten map and looking at the holes, trying to fill them in. Maybe you can't fill them in right away; maybe you have to make a voyage across the sea. It's hard. But you want those details; you want the map to be *complete*, not just "good enough".

Understanding Others

There's a closely related mental stance which I call "ask all the questions". You might think, from the kind of Focusing-like habits I've been describing, that you have to turn within to get the answers. But your focusing object can also be outside of you.

You can orient this toward typical social small-talk. What cognitive habits lead someone to ask questions like "what school did you go to" or "do you have any siblings"? You could have a mental list of standard questions you ask people in social settings. But a different way, which I think is more efficient, is to focus on your "picture" of the person (sort of mentally rehearsing it) and asking questions to fill in the gaps.

Something which surprised me when I tried this attitude on was how self-centred it felt. You're still looking at *your* map for holes. And, you're kind of dominating the conversation, in terms of steering. But, you can bring in the gentle/patient attitude I keep talking about.

You can do the same for topics other than small talk. Maybe you are trying to understand how someone things about X. What many people do is focus mainly on their own picture of X, and let what the other person says kind of land in that map, focusing questions on problems. And that's useful. But you can also focus on *your map of their map*. (This might start out being a copy of your map, since you might assume that they mostly think about X like you and just have some different details. But the cognitive operation is already different; you bring your attention to the places least likely to be the same as for you.)

Again I want to emphasize that I'm talking about a moment-to-moment stance. Not occasionally thinking "what's my map of their map?" during a conversation. Focusing on it *primarily*, letting it drive most of your questions.

This can be a good way of absorbing technical subjects from people.

Hedonic asymmetries

Automatically crossposted

Creating really good outcomes for humanity seems hard. We get bored. If we don't get bored, we still don't like the idea of joy without variety. And joyful experiences only seems good if they are real and meaningful (in some sense we can't easily pin down). And so on.

On the flip side, creating really bad outcomes seems much easier, running into none of the symmetric "problems." So what gives?

I'll argue that nature is basically out to get us, and it's not a coincidence that making things good is so much harder than making them bad.

First: some other explanations

Two common answers (e.g. see here and comments):

- The worst things that can quickly happen to an animal in nature are much worse than the best things that can quickly happen.
- It's easy to kill or maim an animal, but hard to make things go well, so "random" experiences are more likely to be bad than good.

I think both of these are real, but that the consideration in this post is at least as important.

Main argument: reward errors are asymmetric

Suppose that I'm building an RL agent who I want to achieve some goal in the world. I can imagine different kinds of errors:

- **Pessimism**: the rewards are too low. Maybe the agent gets a really low reward even though nothing bad happened.
- **Optimism**: the rewards are too high. Maybe the agent gets a really high reward even though nothing good happened, or gets no reward even though something bad happened.

Pessimistic errors are no big deal. The agent will randomly avoid behaviors that get penalized, but as long as those behaviors are reasonably rare (and aren't the only way to get a good outcome) then that's not too costly.

But *optimistic* errors are catastrophic. The agent will systematically seek out the behaviors that receive the high reward, and will use loopholes to avoid penalties when something actually bad happens. So even if these errors are extremely rare initially, they can totally mess up my agent.

When we try to create suffering by going off distribution, evolution doesn't really care. It didn't build the machinery to be robust.

But when we try to create incredibly good stable outcomes, we are fighting an adversarial game against evolution. Every animal forever has been playing that game using all the tricks it could learn, and evolution has patched every hole that they found.

In order to win this game, evolution can implement general strategies like boredom, or an aversion to meaningless pleasures. Each of these measures makes it harder for us to inadvertently find a loophole that gets us high reward.

Implications

Overall I think this is a relatively optimistic view: some of our asymmetrical intuitions about pleasure and pain may be miscalibrated for a world where we are able to outsmart evolution. I think evolution's tricks just mean that creating good worlds is *difficult* rather than *impossible*, and that we will be able to create an incredibly good world as we become wiser.

It's possible that evolution solved the overoptimism problem in a way that is actually universal—such that it is in fact *impossible* to create outcomes as good as the worst outcomes are bad. But I think that's unlikely. Evolution's solution only needed to be good enough to stop our ancestors from finding loopholes, and we are a much more challenging adversary.

Technology Changes Constraints

A thousand years ago, books were generally written by hand, on parchment made from sheep skin. I don't have a good source on how long it took a person to transcribe a typical book, so for the purpose of this post let's just call it 30 days. I do know that a typical book required the skins of about 12 sheep (source: Braudel).

We can represent this via two production constraints:

$$N_{books} \leq \frac{1}{30} N_{transcriptionDays}$$

$$N_{books} \leq \frac{1}{12} N_{sheep}$$

... and of course we could add more constraints to reflect all the other inputs to a book. We write it like this, rather than just saying "1 book = 12 sheep + 30 transcriptionDays", to highlight that each input is an independent limit on the number of books produced. If we only have 15 sheep on hand, then we can make at most 1 book, no matter how many bored transcriptionists are sitting around.

Another reason why writing out the constraints is useful: it offers a natural way to introduce technology changes.

Let's consider two possible technology changes:

- switching from parchment to paper
- switching from transcriptionists to a printing press

How do these modify the constraints? Well, paper eliminates the sheep constraint and replaces it with a paper constraint (of the form $N_{books} \leq C N_{paper}$ for some C) - yet the transcription constraint remains exactly the same. Conversely, a press eliminates the transcription constraint - yet the sheep constraint remains exactly the same. **The constraint representation is modular with respect to technology changes**: introduction of new technology removes/modifies some constraints, while leaving most of them unaltered.

With a little creativity, this representation can be extended to other kinds of technology changes as well:

- Before the invention of television, we had the constraint $N_{TV} \leq 0$. The invention of television replaced this constraint with a bunch of television production constraints, like $N_{TV} \leq N_{vacuumTubes}$
- Fixed-cost capital goods, e.g. a printing press, add a constraint that we need at least one of the capital good, independent of the number of things produced: $1 \leq N_{press}$
- A more efficient sheep-skin processing technique might replace $N_{books} \leq \frac{1}{12} N_{sheep}$ with $N_{books} \leq \frac{1}{10} N_{sheep}$

... etc.

Conjugacy

One of the main lessons of optimization theory - be it linear programming, convex analysis, what have you - is that every constraint has a conjugate "shadow price" (mathematically given by the Lagrange multiplier). The price indicates how "taut" or "slack" the constraint is - i.e. how much more we can produce if the constraint is relaxed a little bit. If we're a medieval book-maker with 15 sheep and a thousand transcriptionist-years on hand, then the sheep constraint is very taut (more sheep means more books), whereas the transcriptionist constraint is very slack (more transcriptionists does nothing). It's like a rope: pulling on a rope won't do anything unless the rope is taut; hiring more transcriptionists won't do anything unless the transcriptionist constraint is taut. The shadow price quantifies this: it tells us how much the book-maker will pay for additional sheep versus additional transcriptionists. With 15 sheep and a thousand transcriptionist-years, the book-maker will happily pay for more sheep, but will offer roughly zero for more transcriptionists.

Quick recap:

- abundant resource = slack constraint = extra input won't produce much/any extra output <=> producer won't pay for more input = low shadow price of input
- scarce resource = taut constraint = extra input will produce extra output <=> producer will pay for more input = high shadow price of input

If you want to see the math, I highly recommend Stephen Boyd's lectures & book on convex optimization.

So what does all this tell us about technology changes?

Well, new technology removes some constraints and replaces them with new constraints. If the old constraint is slack, then this doesn't do any good. If we already have a million transcriptionist-hours available, and only 15 sheep, then we have no use for a printing press. Consider Pi Cheng: he introduced a movable-type printing press in China around 1045, but it mostly failed to catch on. Why?

Here's one hypothesis: across the board, in many different industries, we see medieval/renaissance China using labor in places where Europe used machines. That suggests that labor, in general, was readily available in China - those constraints were generally slack. Labor in China had a very low shadow price, compared to the shadow price of machinery (i.e. capital goods).

How could we test that hypothesis?

Economic theory provides various conditions under which producers' shadow prices are (roughly) equal to market prices. The simplest such condition is competition, but that assumption usually degrades gracefully: even if competition is less-than-perfect, market prices will still usually *approximate* shadow prices, with the approximation improving as competition increases. So one rough measure of a shadow price is the market price. (Even when the competition assumption fails completely, we can probe the shadow price in other ways - e.g. by looking directly at producers' records, or by looking at how hard producers try to obtain various inputs.)

If competition among book-makers is even remotely reasonable as an approximation, then the book-makers' shadow prices will be close to market prices of the relevant goods. So, we could (very roughly) test our hypothesis by comparing the price of labor relative to capital in medieval/renaissance China to the price of labor relative to capital in medieval/renaissance Europe. Our hypothesis predicts that labor was much cheaper relative to capital in China.

Generalization & Gears

Of course, there are many other possible hypotheses about why movable-type printing didn't catch on in China. A similar approach would apply to other possible hypotheses about the adoption of the press.

For instance, in Europe at least the replacement of parchment with paper is often cited as a key factor, suggesting that *before* the press was adopted, the parchment constraint was much more taut than the transcription constraint - i.e. parchment was a much larger share of the book's price than transcription. Only after the parchment constraint was relaxed did the transcription constraint become more taut, at which point the press caught on.

Note that, in both the capital/labor hypothesis and the paper hypothesis, we don't have a *root* cause. If prices were different, then some upstream factor must have caused the price difference. Rather, constraints/slackness/prices are gears in our model of the world: each constraint/price pair is a gear, which can mediate the causal influence of a wide variety of interventions/root causes.

These gears can interact with each other - the "output" in one constraint may be an "input" in another. For instance, yet another hypothesis for China's non-adoption of movable-type printing is a relative lack of literacy in China; Europe had a much larger market for books. At an economic level, the supply of books was itself a slack constraint in China - people weren't willing to pay much for more books. This constraint would interact with both the capital constraint and the paper constraint - e.g. if few people read books, then there would be little demand for more scalable technologies like printing and paper. Book price/constraint would be a separate gear in the model, alongside the capital, labor, and paper prices/constraints.

Summary

In general, we can reason about the adoption and impact of new technology by looking at prices associated with constraints. If a technology relaxes a slack constraint, then it likely won't be adopted at all, and won't have much impact on total output even if it is adopted. On the other hand, technology which relaxes a taut constraint likely will be adopted, and have a large impact on total output - assuming that the technology doesn't introduce an even more restrictive constraint! (Conversely, though, generalized efficient markets says that new technology which relaxes a taut constraint will be harder to discover in the first place - there was already an incentive to pick the low-hanging fruit.)

Of arguments and wagers

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

Automatically crossposted from ai-alignment.com

(In which I explore an unusual way of combining the two.)

Suppose that Alice and Bob disagree, and both care about Judy's opinion. Perhaps Alice wants to convince Judy that raising the minimum wage is a cost-effective way to fight poverty, and Bob wants to convince Judy that it isn't.

If Judy has the same background knowledge as Alice and Bob, and is willing to spend as much time thinking about the issue as they have, then she can hear all of their arguments and decide for herself whom she believes.

But in many cases Judy will have much less time than Alice or Bob, and is missing a lot of relevant background knowledge. Often Judy can't even understand the key considerations in the argument; how can she hope to arbitrate it?

Wagers

For a warm-up, imagine that Judy could evaluate the arguments if she spent a long enough thinking about them.

To save time, she could make Alice and Bob wager on the result. If both of them believe they'll win the argument, then they should be happy to agree to the deal: "If I win the argument I get $100; if I lose I pay $100." (Note: by the end of the post, no dollars will need to be involved.)

If either side isn't willing to take the bet, then Judy could declare the case settled without wasting her time. If they are both willing to bet, then Judy can hear them out and decide who she agrees with. That person "wins" the argument, and the bet: **Alice and Bob are betting about what Judy will believe, not about the facts on the ground**.

Of course we don't have to stick with 1:1 bets. Judy wants to know the probability that she will be convinced, and so wants to know at what odds the two parties are both willing to bet. Based on that probability, she can decide if she wants to hear the arguments.

It may be that both parties are happy to take 2:1 bets, i.e. each believes they have a 2/3 chance of being right. What should Judy believe? (In fact this should always happen at small stakes: both participants are willing to pay some premium to try to convince Judy. For example, no matter what Alice believes, she would probably be willing to take a bet of $0.10 against $0.01, if doing so would help her convince Judy.)

If this happens, there is an arbitrage opportunity: Judy can make 2:1 bets with both of them, and end up with a guaranteed profit. So we can continuously raise the required stakes for each wager, until either (1) the market approximately clears, i.e. the two

are willing to bet at nearly the same odds, or (2) the arbitrage gap is large enough to compensate Judy for the time of hearing the argument. If (2) happens, then Judy implements the arbitrage and hears the arguments. (In this case Judy gets paid for her time, but the pay is independent of what she decides.)

Recursion

Betting about the whole claim saved us some time (at best). Betting about parts of the claim might get us much further.

In the course of arguing, Alice and Bob will probably rely on intermediate claims or summaries of particular evidence. For example, Alice might provide a short report describing what we should infer from study Z, or Bob might claim “The analysis in study Z is so problematic that we should ignore it.”

Let’s allow anyone to make a claim at any time. But if Alice makes a claim, Bob can make a counterclaim that he feels better represents the evidence. Then we have a recursive argument to decide which version better represents the evidence.

The key idea is that **this recursive argument can also be settled by betting**. So one of two things happens: (1) Judy is told the market-clearing odds, and can use that information to help settle the original argument, or (2) there is an arbitrage opportunity, so Judy hears out the argument and collects the profits to compensate her for the time.

This recursive argument is made in context: that is, Judy evaluates which of the two claims she feels would be a more helpful summary within the original argument. Sometimes this will be a question of fact about which Alice and Bob disagree, but sometimes it will be a more complicated judgment call. For example, we could even have a recursive argument about which wording better reflects the nuances of the situation.

When making this evaluation, Judy uses facts she learned over the course of the argument, but she interprets the claim as she would have interpreted it at the beginning of the argument. For example, if Bob asserts “The ellipsoid algorithm is efficient” and Alice disagrees, Bob cannot win the argument by explaining that “efficient” is a technical term which in context means “polynomial time”—unless that’s how Judy would have understood the statement to start with.

This allows Judy to arbitrate disagreements that are too complex for her to evaluate in their entirety, by showing her what she “would have believed” about a number of intermediate claims, if she had bothered to check. Each of these intermediate claims might itself be too complicated for Judy to evaluate directly—if Judy needed to evaluate it, she would use the same trick again.

Betting with attention

If Alice and Bob are betting about many claims over the course of a long argument, we can replace dollars by “attention points,” which represent Judy’s time thinking about the argument (perhaps 1 attention point = 1 minute of Judy’s time). Judy considers an arbitrage opportunity “good enough” if the profit is more than the time required to evaluate the argument. The initial allocation of attention points reflects the total amount of time Judy is willing to spend thinking about the issue. If someone runs out of attention points, then they can no longer make any claims or use up any of Judy’s time.

This removes some of the problems of using dollars, and introduces a new set of problems. The modified system works best when the total stock of attention points is large compared to the number at stake for each claim. Intuitively, if there are N comparable claims to wager about, the stakes of each should not be more than a 1/sqrt(N) of the total attention pool—or else random chance will be too large a factor. This requirement still allows a large gap between the time actually required to evaluate an argument (i.e. the initial bankroll of attention points) and the total time that would have been required to evaluate all of the claims made in the argument (the total stake of all of the bets). If each claim is itself supported by a recursive argument, this gap can grow exponentially.

Talking it out

If Alice and Bob disagree about a claim (rather, if they disagree about Judy's probability of accepting the claim) then they can have an incentive to "talk it out" rather than bringing the dispute to Judy.

For example, suppose that Alice and Bob each think they have a 60% chance of winning an argument. If they bring in Judy to arbitrate, both of them will get unfavorable odds. Because the surplus from the disagreement is going to Judy, both parties would be happy enough to see their counterparty wise up (and of course both would be happy to wise up themselves). This creates room for positive sum trades.

Rather than bringing in Judy to arbitrate their disagreement, they could do further research, consult an expert, pay Judy attention points to hear her opinion on a key issue, talk to Judy's friends—whatever is the most cost-effective way to resolve the disagreement. Once they have this information, their betting odds can reflect it.

An example

Suppose that Alice and Bob are arguing about how many trees are in North America; both are experts on the topic, but Judy knows nothing about it.

The easiest case is if Alice and Bob know all of the relevant facts, but one of them wants to mislead Judy. In this case, the truth will quickly prevail. Alice and Bob can begin by breaking down the issue into "How many trees are in each of Canada, the US, and Mexico?" If Alice or Bob lie about any of these estimates, they will quickly be corrected. Neither should be willing to bet much for a lie, but if they do, the same thing will happen recursively—the question will be broken down into "how many trees are east and west of the Mississippi?" and so on, until they disagree about how many trees are on a particular hill—a straightforward disagreement to resolve.

In reality, Alice and Bob will have different information about each of these estimates (and geography probably won't be the easiest way to break things down—instead they might combine the different considerations that inform their views, the best guess suggested by different methodologies, approximate counts of each type of tree on each type of land, and so on). If Alice and Bob can reach a rational consensus on a given estimate, then Judy can use that consensus to inform her own view. If Alice and Bob can't resolve their disagreement, then we're back to the previous case. The only difference is that now Alice and Bob have probabilistic disagreements: if Alice disagrees with Bob she doesn't expect to win the ensuing argument with 100% probability, merely with a high probability.

Odds and ends

This writeup leaves many details underspecified. In particular, how does Judy estimate how long it will take her to arbitrate a disagreement? This can be handled in several ways: by having Judy guess, by having Alice and Bob bet on the length of time until Judy reaches a conclusion, by having them make bets of the form “Alice will agree with me with Z effort,” or so on. I don’t know what would work best.

Despite my use of the word “recursion,” the estimate for “time to settle an argument” (which Judy uses to decide when the stakes are high enough to step in and resolve a disagreement) probably shouldn’t include the time required to settle sub-arguments, since Judy is being paid separately for arbitrating each of those. The structure of the arguments and sub-arguments need not be a tree.

This is a simple enough proposal that it can be realistically implemented, so eventually we’ll hopefully see how it works and why it fails.

I expect this will work best if Alice and Bob often argue about similar topics.

This scheme was motivated by a particular exotic application: delegating decision-making to very intelligent machines. In that setting the goal is to scale to very complex disagreements, with very intelligent arguers, while being very efficient with the overseer’s time (and more cavalier with the arguers’ time).

Of arguments and wagers was originally published in AI Alignment on Medium, where people are continuing the conversation by highlighting and responding to this story.

A rant against robots

What comes to your mind when you hear the word "*artificial intelligence*" (or "*artificial general intelligence*")? And if you want to prepare the future, what should come to your mind?

It seems that when most people hear AI, they think of *robots*. Weirdly, this observation includes both laymen and some top academics. Stuart Russell's book (which I greatly enjoyed) is such an example. It often presents robots as an example of an AI.

But this seems problematic to me. I believe that *we should dissociate a lot more AIs from robots*. In fact, given that most people will nevertheless think of robots when we discuss AIs, we might even want to use the terminology *algorithms* rather *AIs*. And perhaps *algorithms with superhuman-level world model and planning capabilities* instead of *AGIs*...

To defend this claim, I shall argue that the most critical aspects of today's and tomorrow's world-scale ethical problems (including x-risks) have and will have to do with algorithms; not robots. Moreover, and most importantly, the example of robots raises both concerns and solutions that seem in fact irrelevant to algorithms. Finally, I'll conclude by arguing that the example of today's large-scale algorithms is actually useful, because it motivates *AI alignment*.

It's about algorithms, not robots!

AIs that matter are algorithms

Today's AI research is mostly driven by non-robotic applications, from natural language processing to image analysis, from protein folding to query answering, from autocompletion to video recommendation. This is where the money is. Google is investing (hundreds of?) millions of dollars in improving its search engine and YouTube's recommendation system. Not in building robots.

Today's ranking, content moderation and automated labeling algorithms are arguably a lot more influential than robots. YouTube's algorithms have arguably become the biggest opinion-maker worldwide. They present risks and opportunities on a massive scale.

And it seems that there is an important probability that tomorrow's most influential algorithms will be somewhat similar, even if they achieve *artificial general intelligence*. Such algorithms will likely be dematerialized, on the cloud, with numerous copies of themselves stored in several data centres and terminals throughout the world.

And they will be extremely powerful. Not because they have some strong physical power. But because they control the flow of information.

The power of information

At the heart of the distinction between algorithms and robots is the distinction between *information* and *matter*. Physics has long turned our attention towards matter and energy. Biology studied on animals, plants and key molecules. Historians focused on monuments, artefacts and industrial revolution. But as these fields grew, they all seem to have been paying more and more attention to information. Physics studied *entropy*. Biology analyzed *gene expressions*. History celebrated the invention of language, writing, printing, and now computing.

Arguably, this is becoming the case for all of our society. Information has become critical to every government, every industry and every charity. Essentially all of today's jobs are actually *information processing* jobs. They are about collecting information, storing information, processing information and emitting information. This very blog post was written after a collection of information, which were then processed and now emitted.

By collecting and analyzing information, you can have a much better idea of what is wrong and what to do. And crucially, by emitting the right information to the right entities, you can start a movement, manipulate individuals and start a revolution. Information is what changes the world.

Better algorithms are information game changers

We humans used to be the leading information processing units on earth. Our human brains were able to collect, store, process and emit information in a way that nothing else on earth could.

But now, there are algorithms. They can collect, store, process and emit far more information than any group of humans ever could. They can now figure out what is wrong and what to do, sometimes far better than we humans can, by learning from information that we humans could not collect, store, process and emit. Algorithms can start movements, manipulate individuals and start revolutions on a global scale. They have become the most powerful entities on earth.

In fact, because such powerful algorithms are deployed by the most powerful companies which also have huge incentives to make their algorithms more capable, it seems much more likely to me that the first *algorithm with superhuman-level world model and planning capabilities* will be much more similar to YouTube's recommendation algorithm than to a robot. Recall that such an algorithm has access to a truly massive amount of data from all over the world. And that data is clearly critical to algorithmic capabilities.

As another example, an algorithm able to send messages through the internet to get a 3D-printer to print killer drones seem a lot more dangerous than any of the killer drones it creates...

This is why I believe that the biggest challenges of AI safety and ethics have likely little to do with robots. These challenges rather seem to concern information and information processing. They are about algorithms.

Not robots. Algorithms.

The case of robots is very misleading

It would be fine to still focus on robots if they were similar enough to algorithms. In the end, I don't really care *why* you would want to solve AI or AGI safety; it just matters (to me) that you do want to solve AI or AGI safety.

Unfortunately though, having a robot in mind as an example of AI or AGI seems also extremely misleading. In fact, so many AGI safety debates could probably be easily shortcut by focusing on algorithms rather than robots.

Distributed algorithms are really hard to interrupt

Let's take the case of safe interruptibility. Many AI safety critics would say that this isn't a problem, because you can just unplug the AI. Well, admittedly, if a robot is not skillful enough to prevent you from unplugging it, and if you have access to its plug, yes, sure, you could probably unplug it.

But now try to unplug an algorithm. Especially a *distributed* algorithm like the YouTube recommendation system or the Google search engine! Even if you were the CEO of Google, I'm skeptical you would be able to interrupt these algorithms.

There's worse. Try to unplug Bitcoin. Well, essentially, you would have to unplug all of the Internet… Good luck with that! This is because Bitcoin was designed to be uninterruptible by any small group of users. This is the whole point of designing distributed algorithms! They are designed to be so-called *Byzantine-fault tolerant*.

It seems more than reasonable to assume that any algorithm with superhuman-level world knowledge and planning capabilities will make sure it is Byzantine-fault tolerant too.

Algorithms work on very different space and time scales

Another key feature of robots that is misleading is that we usually expect them to interact with us at our space and time scale. Intuitively, whatever a robot says can be analyzed. And if what he says is suspicious, we could have the time to correct it before it causes harm.

The case of large-scale algorithms like the YouTube recommendation system is very different. YouTube "speaks" at the rate of millions of recommendations per minute. It "reads" at the rate of 500 hours of videos per minute, and millions of new human behaviours per minute. And YouTube does so on a global scale.

In particular, this means that no human could ever check even a small fraction of what this algorithm does. The mere oversight of large-scale algorithms is way beyond human capability. We need algorithms for algorithmic surveillance.

Today's algorithms already need alignment!

Finally, and perhaps most importantly, robots just aren't here. Even self-driving cars have yet to be commercialized. In this context, it's hard to get people to care about AGI risks, or about alignment. The example of robots is not something familiar to them. It's even associated with science fiction and other futuristic dubious stories.

Conversely, large-scale hugely influential and sophisticated algorithms are already here. And they're already changing the world, with massive unpredictable uncontrollable *side effects*. In fact, it is such *side effects* of algorithm deployments that are *existential risks*, especially if algorithms gain superhuman-level world model and planning capabilities.

Interestingly, today's algorithms also already pose huge ethical problems that absolutely need to be solved. Whenever a user searches "vaccine", "Trump" or "AGI risks" on YouTube, there's an ethical dilemma over which video should be recommended first. Sure, it's not a life or death solution (though "vaccine" could be). But this occurs billions of times per day! And it might make a young scholar mock AGI risks rather than be concerned about them.

Perhaps most interestingly to me, *alignment* (that is, making sure the algorithm's goal is aligned with ours) already seems critical to make today's algorithms *robustly beneficial*. This means that by focusing on the example of today's algorithms, it may be possible to convince AI safety skeptics to do research that is nevertheless useful to AGI safety. As an added bonus, we wouldn't need to sacrifice any respectability.

This is definitely something I'd sign for!

Conclusion

In this post, I briefly shared my frustration to see people discuss AIs and robots often in a same sentence, without clear distinction between the two. I think that this attitude is highly counter-productive to the advocacy of AI risks and the research in AI safety. I believe that we should insist a lot more on the importance of information and information processing through *algorithms*. This seems to me to be a more effective way to promote quality discussion and research on algorithmic alignment.

Review: How to Read a Book (Mortimer Adler, Charles Van Doren)

As part of my research on how to bootstrap understanding in a field, I'm reading books that attempt to answer that question. You might think I should have started with that, but it was useful to get a sense of what problems I needed to solve before I looked for the solution. *How to Read a Book* (affiliate link) is generally very well regarded in this area and came with a strong recommendation from the CEO of Roam, who I would expect to have pretty good thoughts on learning structure. Nonetheless, I was quite disappointed. It took me a long time to put my disappointment into words, but with the help of someone on Facebook I finally figured it out: ***How to Read a Book* is aimed at a narrower subset of books than it acknowledges.** What subset, you might ask? I don't have a great answer, because the authors clearly consider the subset to be the only books, or the only books worth reading, so they didn't leave a lot of clues.

What I can say is that it expects books to follow a rigid structure, and to have a single unifying point (what they call "the unity"). This seems to me to be setting up both the author and the reader to throw out a lot of information because they weren't expecting to see it or couldn't fit it into their existing frameworks- reading like a state, in essence. This is not the only thing that makes me think HtRaB is more about being able to understand *a book* than understand *the world*, although it's the only one I can articulate.

How to Read a Book didn't even attempt to answer my current most important question in reading: How do I know what to save or pay attention to? More attentive reading (including but not limited to note-taking) takes more time and more mental effort. Even if it was free, every additional memory or note eats up space in my brain or exobrain and makes it harder to find other thoughts when I look for them. But I don't necessarily know how important a piece of information is when I read it. Good pre-reading might help me know how important it is to that particular work, but never to my life as a whole.

HtRaB acknowledges that different works have vastly different returns to attention and you should allocate your reading effort accordingly, but I don't feel like it gave me guidance for choosing what to pay attention to, and I have a suspicion that if I pressed the point the authors' answer would be extremely in line with the literary and scientific canon of the 1940s-1970s.

My favorite section of *How to Read a Book* is also the most mechanically detailed: the algorithm for pre-processing a book (explained in detail here). I don't know if this was the most useful to me because it was the most detailed or because a teacher once told me skimming was immoral and I needed that to be challenged.

For the meat of reading, *How to Read a Book* suggests questions to ask but not how to determine the answer. To be fair, this is hard. As I work on my own guide to reading I'm intensely aware of how difficult it is to translate the intentions and external appearance of what I'm doing to inner workings comprehensible to many, or even to other people very similar to myself. I suspect there are people for whom reading these questions causes something to click in their brain and they suddenly start reading better, and that's great, but it makes the book lucky, not good. Which is

nothing to be ashamed of: sometimes a stab at a hard problem is worth more than a perfect solution to an easy one. But HtRaB's stab did not happen to hit my particular problem, nor contain enough deep models to let me make the stab myself.

My overall impression is that this is one of those books that is helpful if you read it at the right time and pretty meh otherwise, and it was the wrong time for me. I also predicted it would be one of those books that's notable for founding a genre but goes on to be surpassed by later books that learned from it, but when I looked at Amazon I found very little. There's lots on speedreading, confusing memorization with learning, and "how to study to pass a test designed by someone else", and I may end up reading some of those because the field is that sparse, but they're not what I actually want. So **if there's a work you or someone you trust likes that attempts to answer any of the following questions, please share it:**

1. How do you find the most likely sources of relevant or useful information?
2. How do you get the most (useful) information out of a sources?
3. How do you decide what information to save?
4. How do you save it in a maximally useful way?

I'm also interested if you have opinions on any of the following:

- Mind Mapping: Improve Memory, Concentration, Communication, Organization, Creativity, and Time Management
    - Kindle Unlimited
- The Lifetime Learner's Guide to Reading and Learning
- Extend Your Mind: Praxis Volume 2
- How to Take Smart Notes
- Accelerated Learning for Expertise: Rapid Knowledge Acquisition Skills to Learn Faster, Comprehend Deeper, and Reach a World-Class Level (Learning how to Learn Book 6) Kindle Edition
- The Self-Learning Blueprint: A Strategic Plan to Break Down Complex Topics, Comprehend Deeply, and Teach Yourself Anything (Learning how to Learn Book 3)
- Writing to Learn
- Unlimited Memory: How to Use Advanced Learning Strategies to Learn Faster, Remember More and be More Productive
- The Art of Reading

So many thanks to my Patreon supporters and the Long Term Future Fund for their support of this research.

Draining the swamp

This is a linkpost for https://rootsofprogress.org/draining-the-swamp

In an earlier post, I outlined our main weapons against infectious disease, including vaccines, antibiotics, antiseptics, pest control, sanitation, and general hygiene. These technologies (in a broad sense, even hand-washing is a technology) have largely eliminated lethal diseases such as smallpox, malaria, cholera, tuberculosis, and polio, at least in the developed world.

But which of these technologies mattered most? Which should we highlight in a history of health and medicine, and which should we hold in our minds as major examples of human progress against disease?

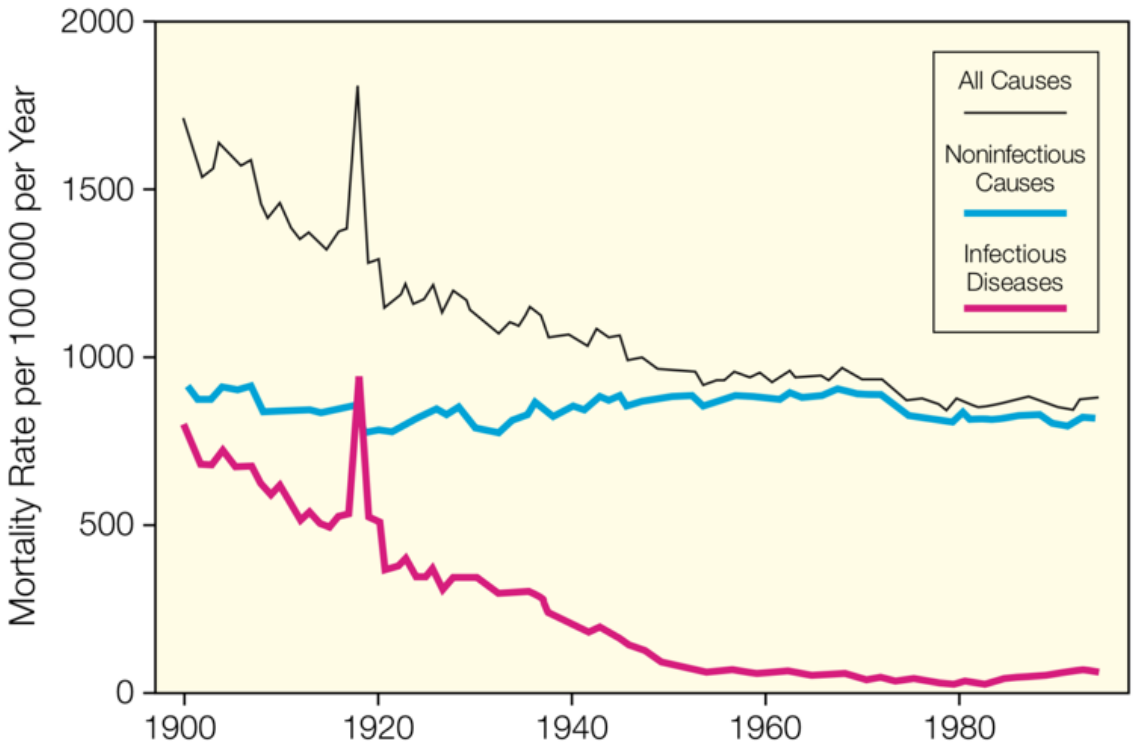
Most histories of medicine give the spotlight to vaccines and antibiotics. They're the most effective medical *treatments*; prior to their introduction, there was little a doctor could do for an infected patient.

But to really answer the question, we should look at mortality rates over time, by disease where possible, and correlate reductions in mortality to specific interventions effective against specific classes of disease. Otherwise we run the risk of assuming that just since something is in all the histories, it must be the most important (leading me to then feature it prominently in *my* history, thereby perpetuating the cycle).

So I started looking into the data and interpretations of it. And the surprising thing I found is that infectious disease mortality rates have been declining steadily since *long* before vaccines or antibiotics.

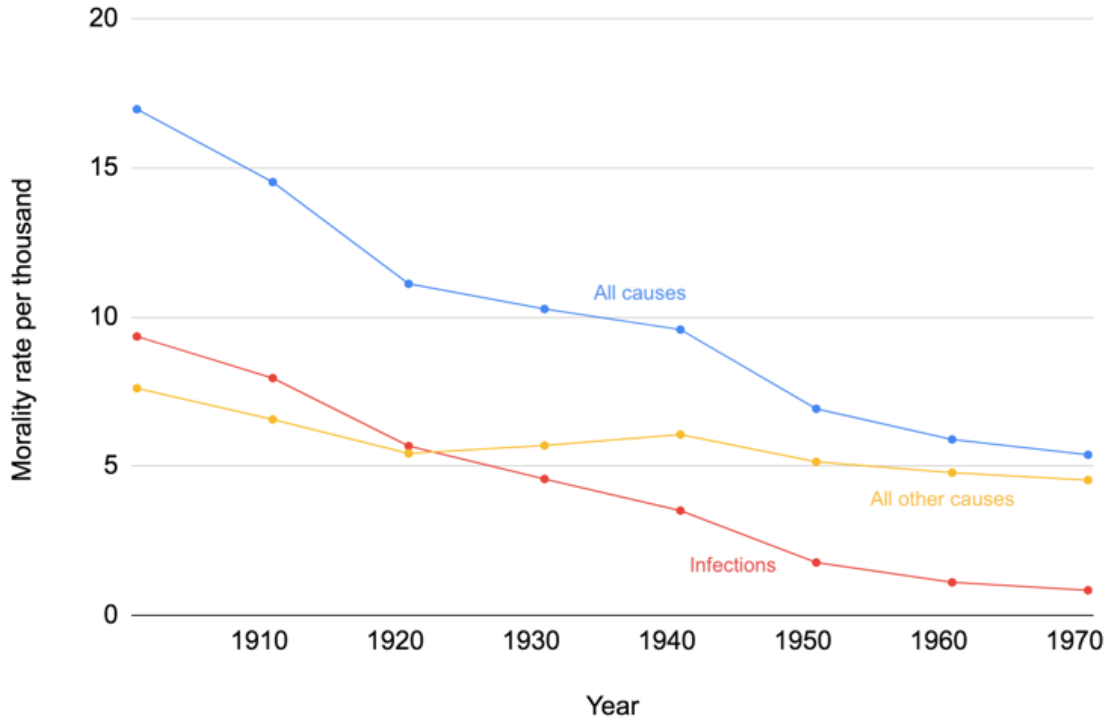
First, let's look at the data that clearly shows that something was going on prior to any effective medical treatment for most diseases.

Here's a chart of 20th century mortality in the US [1] (the spike around 1918 is a major worldwide influenza epidemic, sometimes known as the Spanish Flu):



US crude mortality rates
Armstrong, Conn & Pinner 1999, Fig. 2

And here's a similar chart I made for England and Wales:



Mortality rates in England & Wales, age- and sex-standardized to 1901
Data from McKeown, Record & Turner 1975

Mortality rates in both regions fell, and most of the improvement came from the infectious disease mortality rate, which was reduced by more than an order of magnitude.

Extending the analysis further into the past is difficult. Data is available back to the mid-1800s, but disease classification changes over time as scientific and medical knowledge advances, so it's not always possible to trace the mortality rate from a single disease; and prior to the mid-1800s, no countries that I know of kept reliable cause-of-death records. We do have overall mortality rates from some countries stretching back into the 1800s and even 1700s in some cases, and from these we can see that mortality rates in Europe have been dropping for a long time [2]

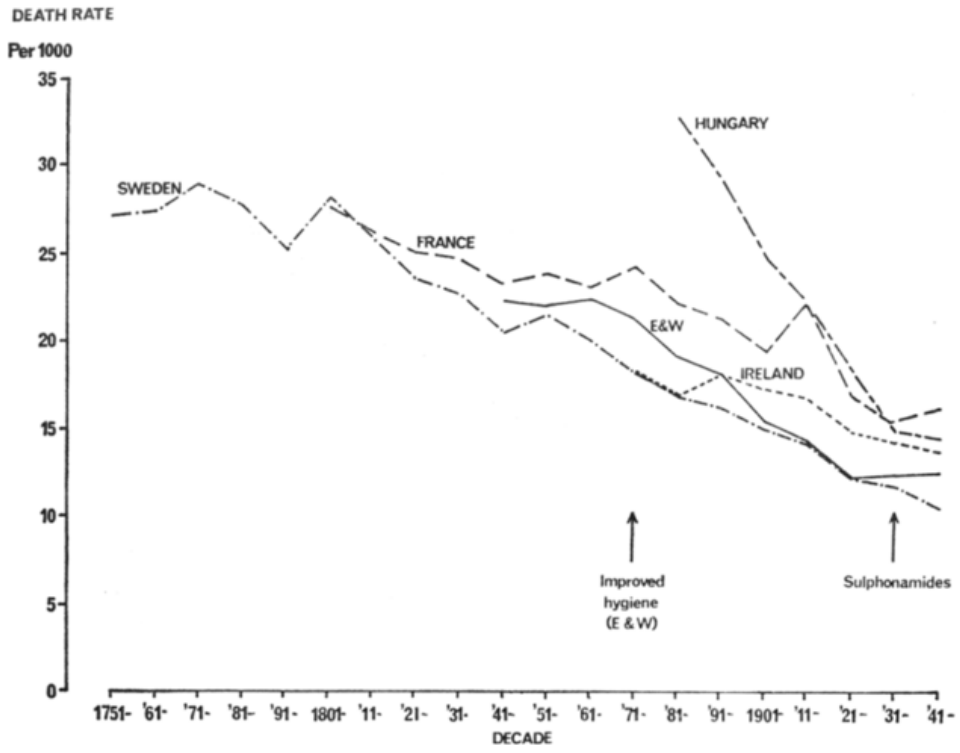


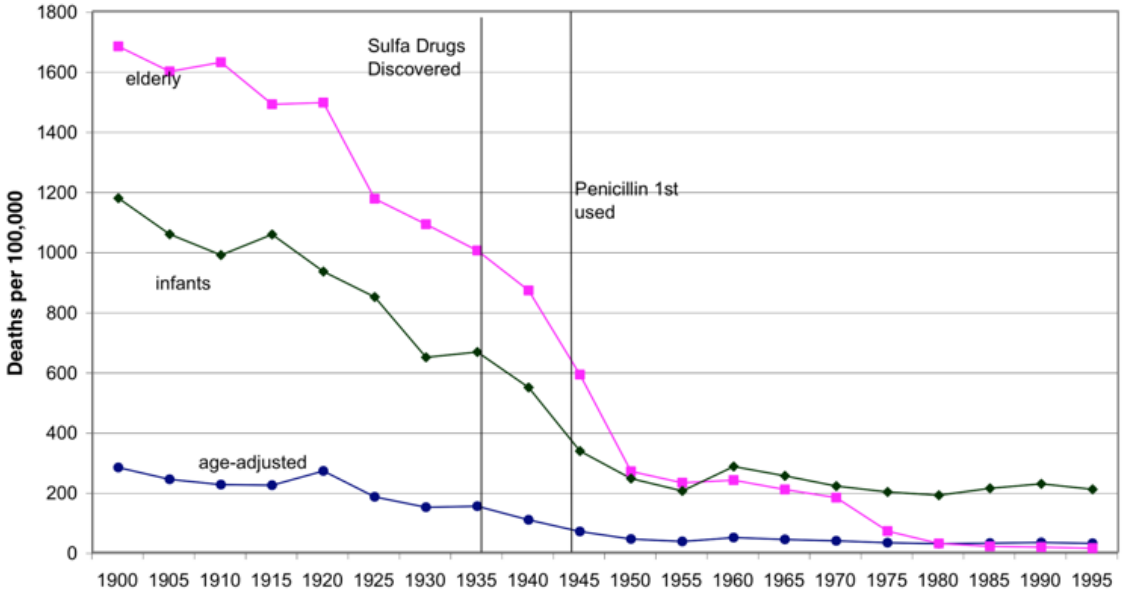
FIGURE 13. *Mean annual death rates of England and Wales, Sweden, France, Ireland, and Hungary from the times when first registered.*

McKeown, Brown & Record 1972

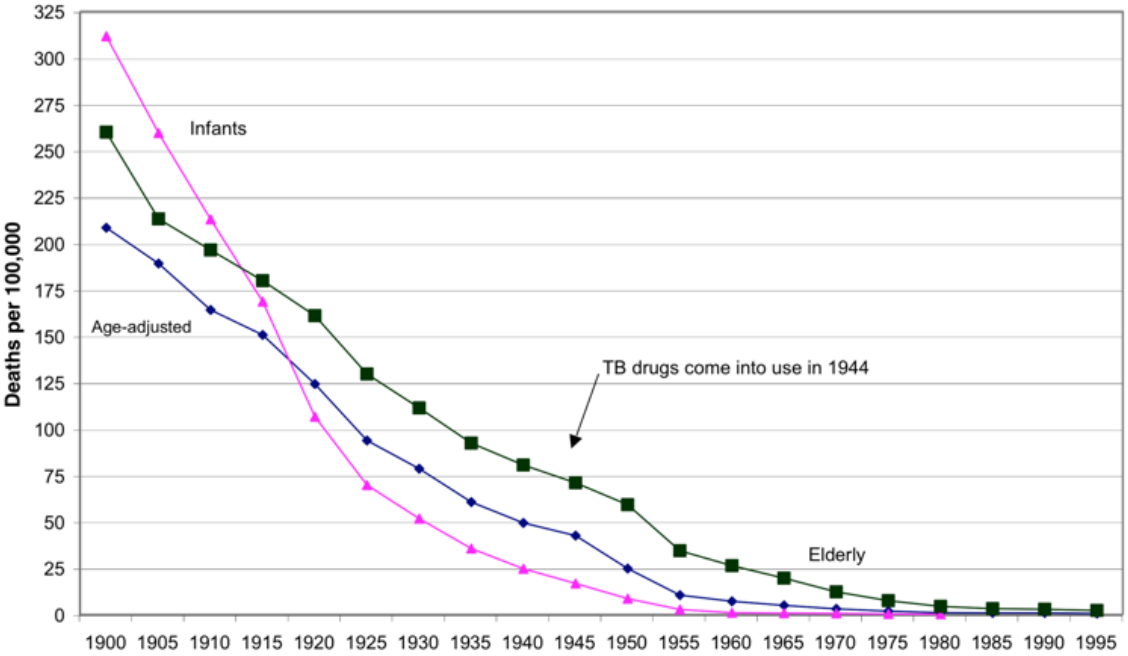
To estimate population and mortality figures prior to these datasets, demographers turn to increasingly limited and unreliable sources, such as parish records of births and deaths, or the London Bills of Mortality, which began running continuously in 1603. These estimates are rough, but they generally show that death rates began to fall in some parts of Europe by 1740 (and in some parts possibly as early as 1670), [3] and that declines in disease mortality were a significant part of this.

In contrast to this timeline, very few effective medical treatments were in widespread use until the late 1930s. Before that time, only a handful of vaccines for major diseases were in use (most notably for smallpox); and there were only a couple of effective pharmaceuticals (most notably for diphtheria, tetanus and syphilis).

The role of some major factor other than medical treatment is even more clear if you break out the mortality rates by specific diseases. In 1900, the most deaths came from tuberculosis, influenza/pneumonia, and gastroenteric diseases such as dysentery. [1][4][5] All of these were effectively conquered by antibiotics in the 1930s and ’40s, but were on the decline since at least the beginning of the century. Some charts that illustrate this in the US: [4]

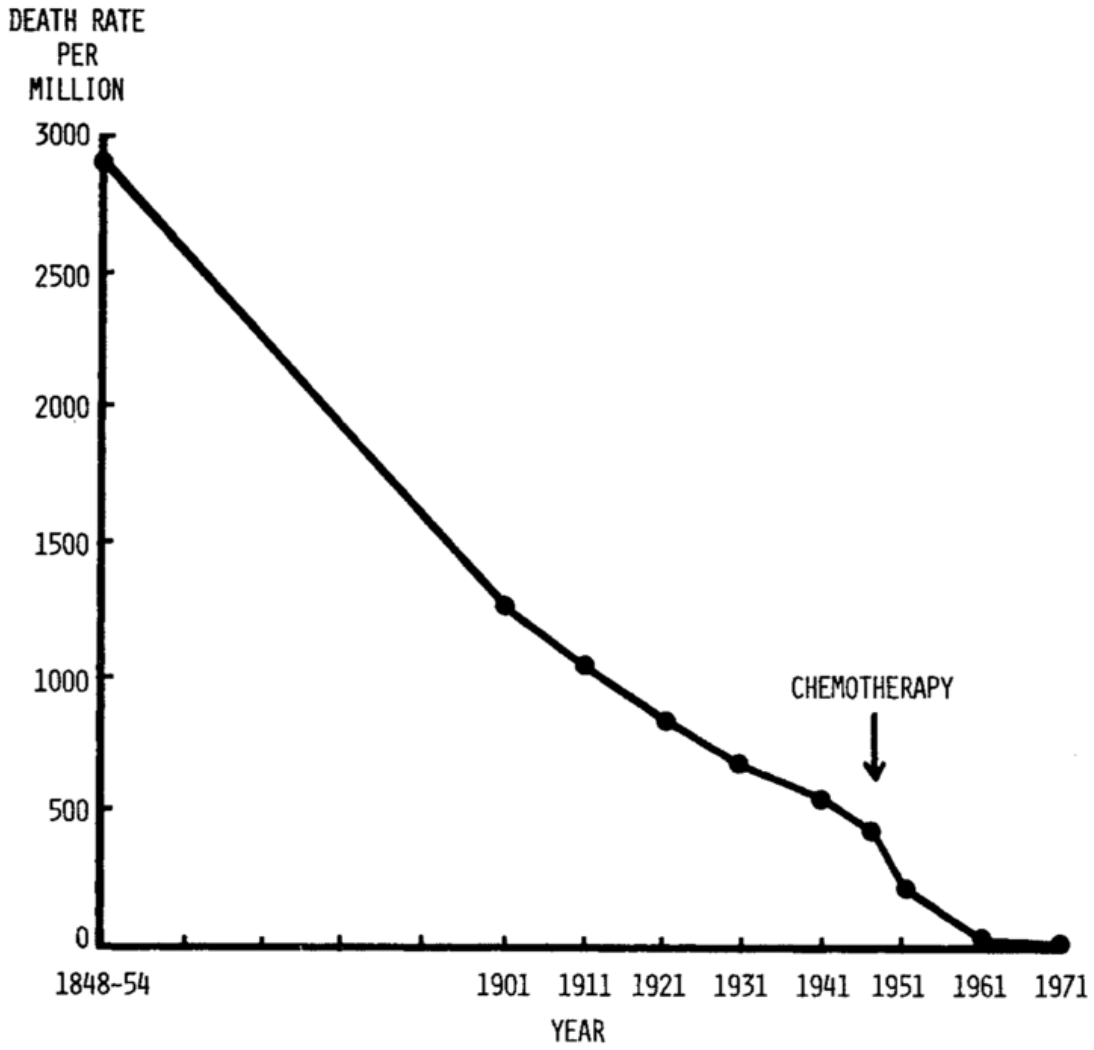


US pneumonia & influenza deaths
Cutler & Meara 2001, Fig. 5



US tuberculosis deaths
Cutler & Meara 2001, Fig. 6

A similar chart for England & Wales [5] (“chemotherapy” here means antibiotics):



Respiratory tuberculosis mortality, standardized to the 1901 age-sex distribution
McKeown, Record & Turner 1975, Fig. 6

Indeed, digging further into the UK data from the late 1800s, we can see that TB was declining since at least 1850 and gastroenteric disease since the 1870s [6]. And similar patterns hold for lesser killers such as measles, which didn't have a vaccine until the 1960s, but which by then had already declined in mortality by more than 90% from its 1900 levels. [1][4]

So what was going on? If you read my survey of technologies against infectious disease, you know that other than drugs and immunization, there is one other way to fight germs: cleaning up the environment.

I was surprised to learn that sanitation efforts began as early as the 1700s—and that these efforts were based on data collection and analysis, long before a full scientific theory of infection had been worked out. James Riley, in "Insects and the European Mortality Decline", writes: [3]

> In the later decades of the seventeenth and early decades of the eighteenth century, a number of internationally renowned physicians … formulated specific measures of

> intervention. Relying on Hippocratic tradition, specifically, on its suggestion that endemic and epidemic diseases are caused by forces in the environment, and influenced by Renaissance efforts at urban sanitation, these physicians proposed to discover the meteorological and topographical forces that might be blamed for the onset of epidemics. Toward this end, they and their followers embarked on a vast campaign to assemble qualitative and quantitative data about epidemics, climate and weather, geographical and topographical signs, and other features of the habitat. Their aim was to find conjunctures or correlations in the data, occasions when epidemics occurred after the same complex of environmental forces. Early signs of such a complex would offer warnings and allow the adoption of measures of prevention and avoidance. This body of medical theory failed to produce a coherent list of correlations, but it did provide a specific body of measures of avoidance and prevention.

In particular, they proposed (each bullet quoted from the article):

- to drain swamps, bogs, moats, and other sites of standing water
- to introduce hydraulic devices that would circulate water in canals and cisterns
- to flush refuse from areas of human habitation
- to ventilate living quarters and meeting places and to burn sulfur sticks or apply other insecticidal measures in houses, hospitals, prisons, meeting halls, and ships
- to inter corpses outside the city
- and by other measures, including refuse burial, to detach humankind from organic waste

These reforms were implemented starting in the 1740s, some by local and central governments, others by “humanitarians acting on private initiative”.

What broad changes were actually implemented, and is it plausible that they had a significant impact?

> To have had a significant effect on insect numbers, the measures proposed by the environmentalists [the physicians advocating environmental cleanup] would have had to have been broadly applied across western Europe. Two measures, lavation and drainage, are particularly important in insect control, and we can focus on examples of their application. Lavation combines programs taking three forms: flushing filth from urban sites, collecting and disposing of refuse, and introducing devices to agitate or circulate standing water. By these means, which would cleanse streets, industrial sites, and buildings, and transform standing water in canals and cisterns into moving water, the environmentalists argued, the city might be made as healthy as the countryside. One model for these proposals was the naturally washed site of the town of Chester, England, where rain periodically flushed refuse into a subterranean drainage network cleansed by tidal action. The objective of the environmentalists was to introduce the same action by hydraulic engineering. Another model was the program followed in Hamburg to collect and dispose of refuse outside the city each day. A third was the improvement of streets by paving and widening, and of urban drainage networks by constructing or expanding sewage systems. Measures of one or another variety were adopted in many British cities and towns in the Improvement Acts of the 1760s and thereafter, and observers, such as William White in York, attributed declines in mortality specifically to them. In Paris, the drainage system was improved in 1740, and later in the century, other measures, including the emptying of cesspits and the installation of sewers, followed. In the Austrian Empire, Johann Peter Frank directed a broad campaign of medical policing, which included projects for refuse collection and disposal.

These efforts affected not only diseases such as malaria, where insects are the primary vector of infection, but also others such as dysentery in which insects (especially flies and cockroaches) can distribute the disease throughout the environment, e.g., from waste to food. Pest control thus provides the best explanation I’ve found for reductions in the mortality rate from the mid-1700s to early 1800s.

After that point, there were some significant shifts in the population and in disease.

As insects became better controlled in the countryside, people began migrating in greater numbers to cities, which worsened disease due to crowding and polluted water. Malaria was on the wane; cholera was on the rise. The mortality rate, after declining for decades, actually plateaued in the mid-1800s, [3] and mortality rates were higher in cities than in the country. [7]

The worsening conditions in the increasingly crowded cities caught the attention of sanitary reformers throughout Europe, such as Edwin Chadwick in Britain and Max von Pettenkofer in Germany, who campaigned in particular for improvements in water and sewage. Starting in the mid-1800s, cities in Europe and the US sought cleaner sources of water, further upstream, or from less polluted rivers. The water was piped into homes, which reduced reliance on wells and surface water. Later, many cities built water filtration systems—the earliest method was simply to allow the water to pass through sand. They also built or modernized sewer systems to transport human waste outside the city and dump it downstream or into the sea, instead of allowing it to collect in pit latrines or cesspools in town, or run through the streets or in open trenches. And crucially, they made sure to keep these systems separate, so that sewage didn't contaminate drinking water.[7][8]

1919 cartoon showing Typhoid as a skeleton monster, beaten to the ground by Filtration Plant and Liquid Chlorine

Sourced from Cutler & Miller 2004

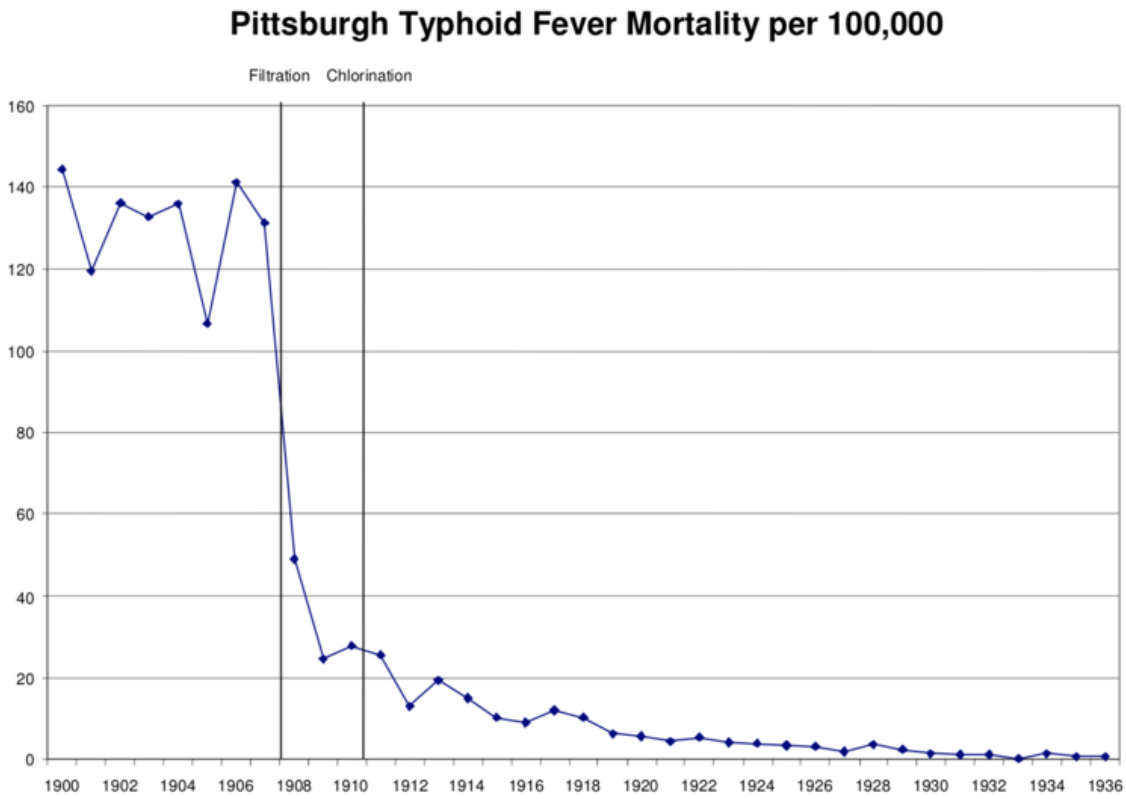
These early efforts, however, were sometimes for aesthetics as much as they were for health, and even to the extent they were aimed at health, they could only be guided by smell, taste and color. [9] But in the late 1800s, the germ theory of disease was established by scientists including Louis Pasteur and Robert Koch. By the 1880s, specific bacteria had been identified as the cause of certain diseases, such as tuberculosis and cholera. Along with this came techniques to grow bacteria in culture and to identify them under the microscope.

These new ideas and methods gave sanitation efforts new tools and targets: instead of aiming to improve sensory qualities, they could aim to eliminate harmful bacteria. Filtration was improved, and chlorine was added to kill germs—first in drinking water, and then in sewage itself. [10]

Cutler & Miller estimate [10] that

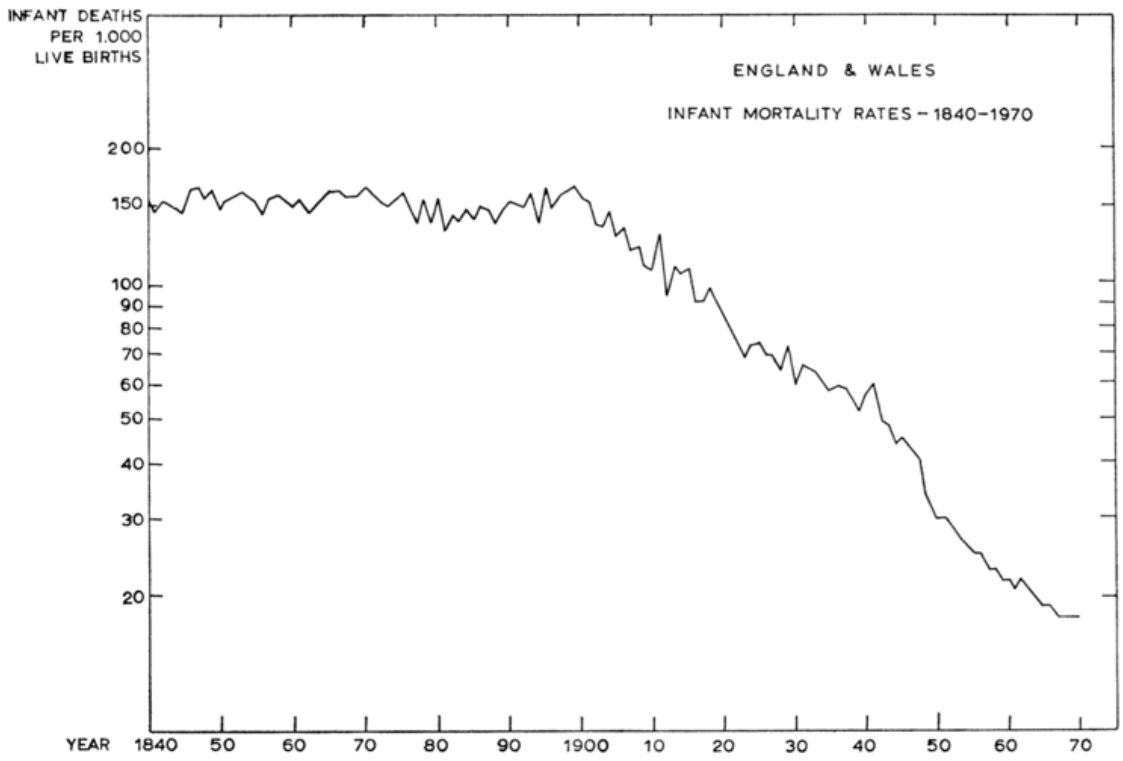
> the introduction of water filtration and chlorination systems led to major reductions in mortality, explaining nearly half of the overall reduction in mortality between 1900 and 1936. Our results also suggest that clean water was responsible for three-quarters of the decline in infant mortality and nearly two-thirds of the decline in child mortality. The magnitude of these effects is striking. Clean water also appears to have led to the near eradication of typhoid fever, a waterborne scourge of the 19th and early 20th Centuries.

They chart mortality from typhoid in several major cities, of which Pittsburgh is the most dramatic:



Cutler & Miller 2004, Fig. 2

Food handling improved as well. Milk is a case in point: In England, as of the late 1800s, milk was transported warm in open containers, making it a literal breeding ground for tuberculosis and other bacterial diseases. Pasteurization was introduced around 1900, along with sealed tins and bottles for transporting and storing milk. Condensed and evaporated milk also became popular around this time, and since these products were sterile they also reduced diseases. All these innovations contributed to the rapid decline in infant mortality seen here: [11]



Beaver 1973, Fig. 1

Finally, the germ theory led to public health efforts to educate the populace on good general hygiene. A policy brief by Samuel Preston says: [9]

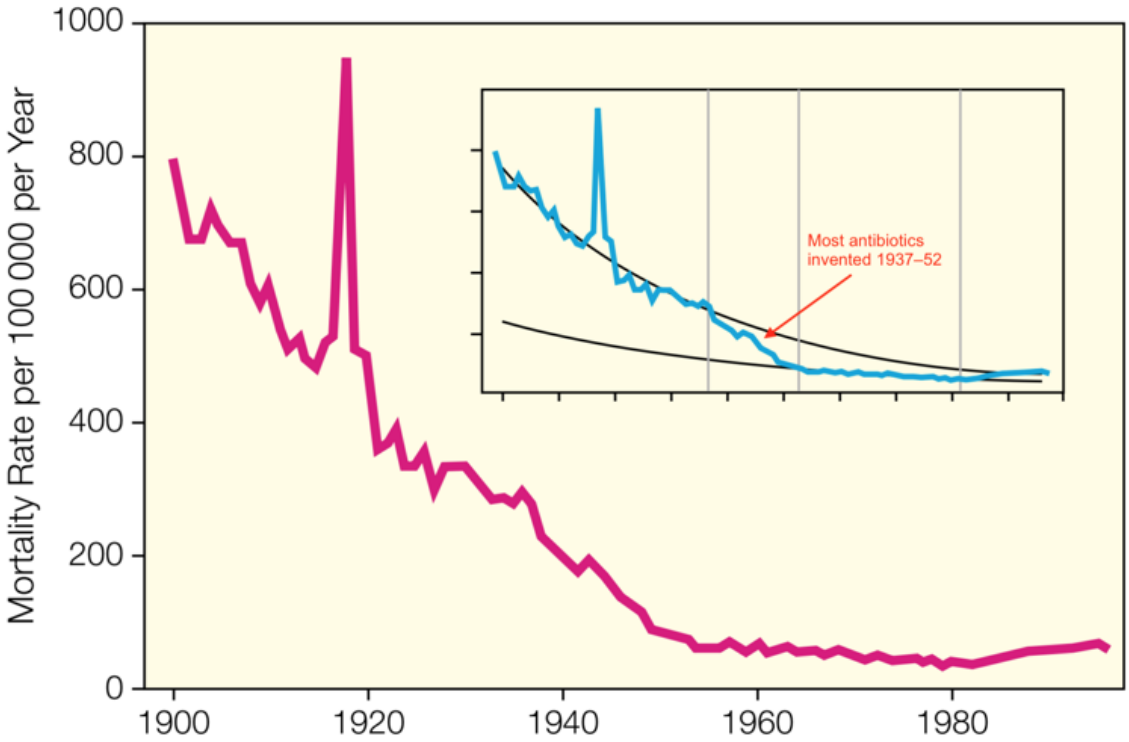
> Enlightened public health officials were quick to recognize how the germ theory should guide their practice. Furthermore, by the time of the first White House Conference on Infant Mortality, held in 1909, they realized that rapid advances in longevity required that public officials go beyond their normal domain of public works and attempt to change the personal health practices of individuals. The germ theory provided a number of powerful weapons for doing so. These included boiling bottles and milk, washing hands, protecting food from flies, isolating sick children, and ventilating rooms. Public health officials launched massive campaigns to encourage these practices. In New York City, milk depots were established with the ostensible purpose of distributing milk to indigent mothers but with the real purpose, according to the director, of instructing mothers in hygienic practices. The New York City Department of Health produced one of the nation's first motion pictures, entitled *The Fly Pest*. At the national level, the new Children's Bureau adopted a primary focus on child health. Its pamphlet called *Infant Care* became the largest selling volume in the history of the Government Printing Office, with some 12 million copies sold by 1940. By the 1920s, the bureau was receiving and answering over 100,000 letters a year from parents seeking child care advice.

Thus the germ theory, long before it led to medical *treatments*, drove down mortality rates by revolutionizing sanitation and hygiene.

So the mortality data points to a large and easy-to-underappreciate role of pest control, water sanitation, food handling, and general hygiene.

However, it also shows the extreme effectiveness of antibiotics, when they were finally invented. In the US, during the golden age of antibiotics from 1937 to '52, the infectious disease mortality rate fell by 8.2%/year, compared to an average of 2.8%/year during 1900–

36 and 2.3%/year during 1953–80. [1] The inset in the following chart illustrates this by overlaying two exponential curves representing the two slower declines before and after; the antibiotics era is in between, where the blue line rapidly falls from the higher curve to the lower:



US crude infectious disease mortality rates
Armstrong, Conn & Pinner 1999, Fig. 1 (annotations added)

Just based on this timing, it's not unreasonable to estimate that antibiotics alone were responsible for a decrease in the mortality rate of something like 5.4%/year for 15 years, or an overall reduction of over 56%.

What about vaccines? Vaccines eradicated smallpox, which was a leading cause of death in the 1700s, although we can't know precisely how much it contributed to 18th-century mortality, since the disease had been greatly diminished by the time causes of death were widely and reliably recorded. A vaccine was also the solution for polio, which caused few deaths compared to many other diseases, but many cases of paralysis. For other diseases, though, vaccines mostly came along late in the game and mopped up what was left after improved sanitation and hygiene had done most of the work.

The other reason that vaccines don't rack up as many points is that, by coincidence, they weren't a good fit for the most common and lethal diseases. The flu has too many strains, and mutates too fast, for a highly effective vaccine; the BCG vaccine for tuberculosis also has varying efficacy (for reasons still not fully understood [12]). Pneumonia and gastroenteric diseases are opportunistic infections that can be caused by any of multiple types of germs; a vaccine only protects against specific germs, and so the vaccines we have for these diseases can only target the most common causes.

However, judged by effectiveness, vaccines score very well, having reduced morbidity for several important diseases by over 99% [13] (see original for notes and caveats):

Disease	Baseline morbidity	1998 morbidity	% decrease
Smallpox	48,164	0	100.0%
Diphtheria	175,885	1	100.0%
Pertussis	147,271	6,279	95.7%
Tetanus	1,314	34	97.4%
Poliomyelitis (paralytic)	16,316	0	100.0%
Measles	503,282	89	100.0%
Mumps	152,209	606	99.6%
Rubella	47,745	345	99.3%
Congenital rubella syndrome	823	5	99.4%
Haemophilus influenzae type b	20,000	54	99.7%

In my previous survey of anti-disease technologies, I also mentioned antiseptics and sterilization. Where do these techniques show up? Primarily in the hospital, it turns out. These techniques are critical for surgical and maternal mortality, reducing post-operation infections and childbed (aka “puerperal”) fever. But diseases of surgery and childbirth are relatively rare compared to those of everyday life, and so they don’t figure prominently in overall mortality rates.

The bottom line is that sanitation—pest control, water filtration and chlorination, safe sewage disposal, milk pasteurization and other food safety, and public education about general hygiene—probably did more than anything else to reduce mortality rates, if only because these techniques were available decades, and in some cases centuries, before anything else. Antibiotics were dramatically effective when they were finally introduced, but by this point a lot of the work had already been done. Vaccines too were extremely effective, but merely delivered the *coup de grace* for many diseases. Other techniques, while very important in limited spheres, simply addressed problems that were too small to show up on any of the top lists.

What strikes me about all this is a similar pattern to what I’ve previously written on science & the Industrial Revolution:

1. A naive or cursory look at the history gives a simplistic account: Medicine reduced disease! Science saves lives!
2. A closer look reveals that disease mortality was dropping long before antibiotics or vaccines. So (some hastily conclude) medicine didn’t really matter after all—so much for better living through science!
3. An even closer look shows that actually, the germ theory led to sanitation and hygiene improvements decades before we had specific treatments. So, as with the steam engine, it turns out science *was* relevant, just not in the obvious first place one might look.
4. Finally, the galaxy-brain take looks not only at direct influences but indirect/cultural ones: The Scientific Revolution led to new ways of experimenting and collecting/analyzing data that led to practical improvements (in waste disposal and insect control) long before we had a fundamental scientific theory.

Thanks to Tyler Cowen, Matt Bateman, Andrew Layman, Sean Pawley, Ben Landau-Taylor, and Michael Goff for reviewing drafts of this post.

References

1. Armstrong, G. L., Conn, L. A. & Pinner, R. W. Trends in Infectious Disease Mortality in the United States During the 20th Century. *JAMA* **281**, 61–66 (1999).
2. McKeown, T., Brown, R. G. & Record, R. G. An interpretation of the modern rise of population in Europe. *Population Studies* **26**, 345–382 (1972).
3. Riley, J. C. Insects and the European Mortality Decline. *The American Historical Review* **91**, 833–858 (1986).
4. Cutler, D. & Meara, E. *Changes in the Age Distribution of Mortality Over the 20th Century*. http://www.nber.org/papers/w8556 (2001) doi:10.3386/w8556.
5. McKeown, T., Record, R. G. & Turner, R. D. An interpretation of the decline of mortality in England and Wales during the twentieth century. *Population Studies* **29**, 391–422 (1975).
6. McKeown, T. & Record, R. G. Reasons for the decline of mortality in england and wales during the nineteenth century. *Population Studies* **16**, 94–122 (1962).
7. Szreter, S. The Importance of Social Intervention in Britain's Mortality Decline c.1850–1914: a Re-interpretation of the Role of Public Health1. *Social History of Medicine* **1**, 1–38 (1988).
8. Burström, B., Macassa, G., Öberg, L., Bernhardt, E. & Smedman, L. Equitable Child Health Interventions. *American Journal of Public Health* **95**, 208–216 (2005).
9. Preston, S. H. American Longevity: Past, Present, and Future. (1996) doi:10.2139/ssrn.1824586.
10. Cutler, D. M. & Miller, G. *The Role of Public Health Improvements in Health Advances: The 20th Century United States*. http://www.nber.org/papers/w10511 (2004) doi:10.3386/w10511.
11. Beaver, M. W. Population, Infant Mortality and Milk. *Population Studies* **27**, 243–254 (1973).
12. Dockrell, H. M. & Smith, S. G. What Have We Learnt about BCG Vaccination in the Last 20 Years? *Frontiers in Immunology* **8**, 1134 (2017).
13. CDC & Bonanni, P. Achievements in Public Health, 1900–1999: Impact of Vaccines Universally Recommended for Children–United States, 1990–1998. *Demographic Impact of Vaccination: A Review* **48**, 243–248 (1999).

Exploring safe exploration

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

This post is an attempt at reformulating some of the points I wanted to make in "Safe exploration and corrigibility" in a clearer way. This post is standalone and does not assume that post as background.

In a previous comment thread, Rohin argued that safe exploration is currently defined as being about the agent not making "an accidental mistake." I think that definition is wrong, at least to the extent that I think it both doesn't make much sense and doesn't describe how I actually expect current safe exploration work to be useful.

First, what does it mean for a failure to be an "accident?" This question is simple from the perspective of an engineer outside the whole system—any unintended failure is an accident, encapsulating the majority of AI safety concerns (i.e. "accident risk"). But that's clearly not what the term "accidental mistake" is pointing at in this context—rather, the question here is *what is an accident from the perspective of the model?* Intuitively, an accident from the perspective of the model should be some failure that the model didn't intend or wouldn't retroactively endorse. But that sort of a definition only makes sense for highly coherent mesa-optimizers that actually have some notion of intent. Maybe instead we should be thinking of this from the perspective of the base optimizer/loss function? That is, maybe a failure is an accidental failure if the loss function wouldn't retroactively endorse it (e.g. the model got a very low reward for making the mistake). By this definition, however, *every generalization failure is an accidental failure* such that safe exploration would just be the problem of generalization.

Of all of these definitions, the definition defining an accidental failure from the perspective of the model as a failure that the model didn't intend or wouldn't endorse seems the most sensical to me. Even assuming that your model is a highly coherent mesa-optimizer such that this definition makes sense, however, I still don't think it describes current safe exploration work, and in fact I don't think it's even really a safety problem. The problem of producing models which don't make mistakes from the perspective of their own internal goals is precisely the problem of making powerful, capable models—that is, it's precisely the problem of capability generalization. Thus, to the extent that it's reasonable to say this for any ML problem, the problem of accidental mistakes under this definition is just a capabilities problem. However, I don't think that at all invalidates the utility of current safe exploration work, as I don't think that current safe exploration work is actually best understood as avoiding "accidental mistakes."

If safe exploration work isn't about avoiding accidental mistakes, however, then what is it about? Well, let's take a look at an example. Safety Gym has a variety of different environments containing both goal states that the agent is supposed to reach and unsafe states that the agent is supposed to avoid. From OpenAI's blog post: "If deep reinforcement learning is applied to the real world, whether in robotics or internet-based tasks, it will be important to have algorithms that are safe even while learning—like a self-driving car that can learn to avoid accidents without actually having to experience them." Why wouldn't this happen naturally, though—shouldn't an agent in

a POMDP always want to be careful? Well, not quite. When we do RL, there are really two different forms of exploration happening:[1]

- **Within-episode exploration,** where the agent tries to identify what particular environment/state it's in, and
- **Across-episode exploration,** which is the problem of making your agent explore enough to gather all the data necessary to train it properly.

In your standard episodic POMDP setting, you get within-episode exploration naturally, but not across-episode exploration, which you have to explicitly incentivize.[2] Because we have to explicitly incentivize across-episode exploration, however, it can often lead to behaviors which are contrary to the goal of actually trying to achieve the greatest possible reward in the current episode. Fundamentally, I think current safe exploration research is about trying to fix that problem—that is, it's about trying to make across-episode exploration less detrimental to reward acquisition. This sort of a problem is most important in an online learning setting where bad across-episode exploration could lead to catastrophic consequences (e.g. crashing an actual car to get more data about car crashes).

Thus, rather than define safe exploration as "avoiding accidental mistakes," I think the right definition is something more like "improving across-episode exploration." However, I think that this framing makes clear that there are other types of safe exploration problems—that is, there are other problems in the general domain of making across-episode exploration better. For example, I would love to see an exploration of how different across-episode exploration techniques impact capability generalization vs. objective generalization—that is, when is across-episode exploration helping you collect data which improves the model's ability to achieve its current goal versus helping you collect data which improves the model's goal?[3] Because across-episode exploration is explicitly incentivized, it seems entirely possible to me that we'll end up getting the incentives wrong somehow, so it seems quite important to me to think about how to get them right—and I think that the problem of getting them right is the right way to think about safe exploration.

---

1. This terminology is borrowed from Rohin's first comment in the same comment chain I mentioned previously. ↩

2. With some caveats—in fact, I think a form of across-episode exploration will be instrumentally incentivized for an agent that is aware of the training process it resides in, though that's a bit of a tricky question that I won't try to fully address now (I tried talking about this somewhat in "Safe exploration and corrigibility," though I don't think I really succeeded there). ↩

3. This is what I somewhat confusingly called the "objective exploration problem" in "Safe exploration and corrigibility." ↩

Why Do You Keep Having This Problem?

One thing I've noticed recently is that when someone complains about how a certain issue "just keeps happening" or they "keep having to deal with it", it often seems to indicate an unsolved problem that people may not be aware of. Some examples:

- Players of a game repeatedly ask the same rules questions to the judges at an event. This doesn't mean everyone is bad at reading -- it likely indicates an area of the rules that is unclear or misleadingly written.
- People keep trying to open a door the wrong way, either pulling on a door that's supposed to be pushed or pushing a door that's supposed to be pulled -- it's quite possible the handle has been designed poorly in a way that gives people the wrong idea of how to use it. (The Design of Everyday Things has more examples of this sort of issue.)
- Someone keeps hearing the same type of complaint or having the same conversation about a particular policy at work -- this might be a sign that that policy might have issues. [1]
- Every time someone tries to moderate a forum they run, lots of users protest against their actions and call it unjust; this might be a sign that they're making bad moderation decisions.

I'm not going to say that *all* such cases are ones where things should change -- it's certainly possible that one might have to take unpopular but necessary measures under some circumstances -- but I do think that this sort of thing should be a pretty clear warning sign that things might be going wrong.

Thus, I suspect you should consider these sorts of patterns not just as "some funny thing that keeps happening" or whatever, but rather as potential indicators of "bugs" to be corrected!

[1] This post was primarily inspired by a situation in which I saw someone write "This is the fifth time I've had this conversation in the last 24 hours and I'm sick of it" or words to that effect -- the reason they had kept having that conversation, at least in my view, was because they were implementing a bad policy and people kept questioning them on it (with perhaps varying degrees of politeness).

The Road to Mazedom

Previous post: How Escape From Immoral Mazes

Sequence begins here: Moloch Hasn't Won

The previous posts mostly took mazes as given.

As an individual, one's ability to fight any large system is limited.

That does not mean our individual decisions do not matter. They do matter. They add up.

Mostly our choice is a basic one. Lend our strength to that which we wish to be free from. Or not do so.

Even that is difficult. The methods of doing so are unclear. Mazes are ubiquitous. Not lending our strength to mazes, together with the goal of keeping one's metaphorical soul intact and still putting food on the table, is already an ambitious set of goals for an individual in a world of mazes.

We now shift perspective from the individual to the system as a whole. We stop taking mazes as given.

It is time to ask *why* and *how* all of this happens, and what if anything we can do, individually or collectively, about it.

In particular, this post presents an explicit model of the first question, which is:

How did mazes come to be, both individually and overall?

This is partly a summary of the model developed so far. It is partly making the model more explicit, and partly the fleshing out of that model with more gears.

If there are points here for which you believe the previous posts failed to lay the groundwork they should have laid, please note that in the comments by number, so I can consider fixing that. Keep this distinct from any disagreements or other notes on these claims, which are also welcome.

1. Every organization has an organizational culture. That culture can and does change.
2. Those who focus on their own advancement at the expense of other considerations will, by default, advance further, faster and more often. Those who do not do this will not advance. Increasing amounts of focus make this effect increasingly large.
3. Focus on one's own advancement inside hierarchies causes individuals to self-modify in order to be the type of person who automatically engages in maze-creating and maze-supporting behaviors. They will also see such behavior as natural and virtuous.
4. Middle management performance is inherently difficult to assess. Maze behaviors systematically compound this problem. They strip away points of differentiation beyond loyalty to the maze and willingness to sacrifice one's self on its behalf, plus politics. Information and records are destroyed. Belief in the

possibility of differentiation in skill level, or of object-level value creation, is destroyed.

5. The more one is already within a maze, the more one is rewarded for maze-creating and maze-supporting behaviors, and for self-modifications towards such behaviors. This creates a vicious cycle.
6. Focus on one's own advancement causes one to wish to ally with others who do the same thing. That means allying with those who are engaging in maze-creating behaviors, and who are likely to do so in the future. Those people are likely to have future power. They are aligned in support of your new values and likely actions.
7. Changing the organizational culture towards a maze, which we will call *raising the maze level,* benefits those who wish to engage in maze-like behavior at the expense of those who do not. Those wishing to raise maize levels implicitly, and sometimes explicitly, coordinate together to reward maze behaviors, culture and values, and punish other behaviors, cultures and values.
8. Those who wish to have strong allies notice that strong potential allies want to ally with those who support mazes, and with those who ally with those who support mazes, and so on. This creates a strong incentive to strongly signal, in a way that others who support mazes can recognize, our support for maze behaviors and rising maze levels. One does this by supporting maze behaviors and maze allies whenever possible over all other considerations, without any need for explicit coordination or reciprocity, and by other costly signals of maze virtue.
9. The larger an organization and the more of a maze it becomes, the closer competition among its middle managers resembles super-perfect competition plus political considerations. Slack is destroyed. Those who refuse to get with the program, where the essence of the program *is support of mazes over non-mazes,* stop getting promoted or are pushed out entirely.
10. Such behaviors chose how to react to people largely by observing those people's culture and values. Those who wish to get ahead in such worlds must self-modify to instinctively support such actions, whether or not doing so is locally in their self interest. Being too aware of one's local self-interest is therefore not in one's broader self-interest. Humans are much better at doing all this, and at detecting it, than they are at faking it. The way one gets such behavior is through cultivating habits, including habits of thought, and choosing one's virtues. This self-modification creates even stronger implicit coordination.
11. There are contravening forces that can potentially outweigh all these effects, and result in maze behaviors being net punished. But they require those opposed to maze behaviors, culture and values to devote substantial resources to the cause, and to bear substantial costs. The more of a maze a place has already become, the harder it will be to turn things around or even stop things from getting worse. If such efforts are to succeed, this needs to be a high priority.
12. Even if maze behaviors are net punished for now, those who have embraced the maze nature will be skeptical of this. Even if they observe such behaviors being net punished now, they will not expect this to continue. Given the state of our world and culture, this is a highly reasonable prior. They also have knowledge of their own maze nature and presence within the organization, which is likely to raise maze levels over time and is evidence that the fight against mazedom is failing. And they stand to gain a huge competitive edge by raising the local maze level. Thus, the fight never ends.
13. Damage done is very difficult to reverse. Once particular maze behaviors become tolerated and levels rise, it takes a lot of effort to undo that.

14. Once people who support mazes are in places of authority in a given area, that area will rapidly become a maze. This is true of organizations, and of sub-organizations with an organization.
15. If the head of an organization believes in mazes, and has the time to choose and reward the people of their choice, it's all over. Probably permanently.
16. Mazes reward individuals who engage in maze behaviors and exhibit maze culture and values, and punish those who do not so exhibit, even outside the maze or organization in question. This includes customers, producers, business partners, investors and venture capitalists, board members, analysts, media, government officials, academics and anyone else who supports or opposes such patterns.
17. All strengthening of mazes anywhere creates additional force supporting mazes elsewhere. Mazes instinctively support other mazes. As society falls increasingly under the sway of mazes, it implicitly cooperates to push everyone and everything into supporting the behaviors, culture and values of mazes.
18. The end result inside any given organization is that maze behaviors grow stronger and more common over time. This is balanced by maze behaviors making the organization less effective, and thus more likely to fail.
19. Occasionally an organization can successfully lower its maze level and change its culture, but this is expensive and rare heroic behavior. Usually this requires a bold leader and getting rid of a lot of people, and the old organization is effectively replaced with a new one, even if the name does not change. A similar house cleaning happens more naturally in the other direction when and as maze levels rise.
20. Maze behaviors grow stronger and more common over time in any given organization barring rare heroic efforts. As organizations get bigger and last longer, maze levels increase.
21. When interacting with a world of low maze levels, or especially when interacting with individuals who have not embraced the maze nature, mazes are at a large competitive disadvantage versus non-mazes. Organizations with too-high maze levels become more likely to fail.
22. As organizations fail and are replaced by smaller upstarts via creative destruction, revolution or other replacement, maze levels decrease.
23. Replacement of old organizations by new ones is the primary way maze levels decline.
24. As the overall maze level rises, mazes gain a competitive advantage over non-mazes.
25. If society sufficiently rewards mazes and punishes non-mazes, non-mazes can stop failing less often than mazes. Existing organizations become increasingly propped up by corruption. New organizations will start off increasingly maze-like, signal their intent to become mazes, and raise their maze levels more rapidly. They will still usually start out at much lower maze levels than old organizations.
26. New organizations and smaller organizations also have more benefit in survival and growth from non-maze behaviors versus maze behaviors, as they have a greater need to do things mazes cannot do, or that they cannot do without huge additional overhead. Even in the scenario where large organizations benefit from maze coordination more than they are hurt by maze inefficiency, it can still benefit smaller organizations to minimize maze levels.
27. Mazes have reason to and do obscure that they are mazes, and to obscure the nature of mazes and maze behaviors. This allows them to avoid being attacked or shunned by those who retain enough conventional not-reversed values that they would recoil in horror from such behaviors if they understood them, and potentially fight back against mazes or to lower maze levels. The maze embracing individuals also take advantage of those who do not know of the

maze nature. It is easy to see why the organizations described in *Moral Mazes* would prefer people not read the book *Moral Mazes.*

28. Simultaneously with pretending to the outside not to be mazes, those within them will claim if challenged that everybody knows they are mazes and how mazes work.
29. As maze levels rise, mazes take control of more and more of an economy and people's lives.
30. Under sufficiently strong pressure the maze behaviors, value and culture filter out into the broader society. Maze behaviors, values and culture are seen increasingly as legitimate and comfortable and praiseworthy. This happens even outside of any organization. Non-maze behaviors are increasingly seen as illegitimate, uncomfortable and blameworthy.
31. The result of these effects is that people in societies with high maze levels, especially those with power and wealth, increasingly and increasingly openly oppose and vilify the creation of clarity, engaging in any productive object-level action, and participation in or even belief in the existence of positive sum games of any kind. Simulacrum levels continue to rise.
32. Given sufficiently high societal maze levels, talk in support of maze behaviors would eventually becoming more and more open, and dominate discourse and how people are educated about the world, as people explicitly and publicly endorse and teach anti-virtues over virtues.
33. We would see increasing societal inability to create clarity, engage in actions or do anything other than repeat existing patterns. Costs to all still possible actions would rise. Existing patterns would expropriate increasing portions of remaining resources to keep themselves afloat, and increasingly ban any activity outside those patterns.
34. The default outcome on the scale of individual organizations is the rise and fall of those organizations over time.
35. The default short term outcome on the scale of a nation, when maze levels and simulacrum levels increase, is declining growth, dynamism, slack, discourse, hope, happiness, virtue and wealth. People increasingly lose the things that matter in life.
36. The default long term outcome on the scale of a nation is the rise and fall of civilizations.
37. We do in fact see all of this. Here and now.

The next few posts will flesh this out more and provide, as best I can, answers to the other questions.

Circling as Cousin to Rationality

Often, I talk to people who are highly skeptical, systematic thinkers who are frustrated with the level of *inexplicable* interest in Circling (previously discussed on LW) among some rationalists. “Sure,” they might say, “I can see how it might be a fun experience for some people, but why give it all this attention?” When people who are interested in Circling can’t give them a good response besides “try it, and perhaps then you’ll get why we like it,” there’s nothing in that response that distinguishes a contagious mind-virus from something useful for reasons not yet understood.

This post isn’t an attempt to fully explain what Circling is, nor do I think I’ll be able to capture everything that’s good about Circling. The hope is to clearly identify one way in which Circling is deeply principled in a way that rhymes with rationality, and potentially explains a substantial fraction of rationalist interest in Circling. As some context; I’m certified to lead Circles in the Circling Europe style after going through their training program, but I’ve done less Circling than Unreal had when she wrote this post, and I have minimal experience with the other styles.

Why am I interested in Circling?

Fundamentally, I think the thing that sets Circling apart is that it focuses on updating based on experience and strives to create a tight, high-bandwidth feedback loop to generate that experience. Add in some other principles and reflection, and you have a functioning culture of empiricism directed at human connection and psychology. I think they’d describe it a bit differently and put the emphasis in different places, while thinking that my characterization isn’t too unfair. This foundation of empiricism makes Circling seem to me like a ‘cousin of Rationality,’ though focused on people instead of systems.

I first noticed the way in which Circling was trying to implement empiricism early in my Circling experience, but it fully crystallized when a Circler said something that rhymes with P.C. Hodgell’s “That which can be destroyed by the truth should be.” I can’t remember the words precisely, but it was something like “in the practice, I have a deep level of trust that I should be open to the universe.” That is, he didn’t trust that authentic expression will predictably lead to success according to his current goals, but rather that a methodological commitment to putting himself out there and seeing what happens would lead to deeper understanding and connection with others, even though it requires relinquishing attachment to specific goals. This is a cognitive clone of how scientists don’t trust that running experiments will predictably lead to confirmation of their current hypotheses, but rather that a methodological commitment to experimentation and seeing what happens would lead to a deeper understanding of nature. A commitment to natural science is fueled by a belief that the process of openness and updating is worth doing; a commitment to human science is fueled by a belief that the process of openness and updating is worth doing.

Why should “that which can be destroyed by the truth” be destroyed? Because the truth is fundamentally more real and valuable than what it replaces, which must be implemented on a deeper level than “what my current beliefs think.” Similarly, why should “that which can be destroyed by authenticity” be destroyed? Because authenticity [IOU: a link as good as 'The Simple Truth'] is fundamentally more real and

valuable than what it replaces, which must be implemented on a deeper level than "what my current beliefs think." I don't mean to pitch 'radical honesty' here, or other sorts of excessive openness; authentic relationships include distance and walls and politeness and flexible preferences.

What is Circling, in this view?

So what is Circling, and why do I think it's empirical in this way? I sometimes describe Circling as "multiplayer meditation." That is, like a meditative practice, it involves a significant chunk of time devoted to attending to your own attention. Unlike sitting meditation, it happens *in connection* with other people, which allows you to see the parts of your mind that activate around other people, instead of just the parts that activate when you're sitting with yourself. It also lets you attend to what's happening in other people, both to get to understand them better and to see the ways in which they are or aren't a mirror of what's going on in you. It's sometimes like 'the group' trying to meditate about 'itself.' A basic kind of Circle holds one of the members as the 'object of meditation', like a mantra or breathing with a sitting meditation, with a different member acting as facilitator, keeping the timebox, opening and closing, and helping guide attention towards the object when it drifts. Other Circles have no predefined object, and go wherever the group's attention takes them.

As part of this exploration, people often run into situations where they don't have social scripts. Circling has its own set of scripts that allow for navigation of trickier territory, and also trains script-writing skills. They often run into situations that are vulnerable, where people are encouraged to follow their attention and name their dilemmas; if you're trying to deepen your understanding of yourself and become attuned to subtler distinctions between experiences and emotions, running roughshod over your boundaries or switching them off is a clumsy and mistaken way to do so. Circles often find themselves meditating on why they cannot go deeper in that moment, not yet at least, in a way that welcomes and incorporates the resistance.

Circling Europe has five principles; each of these has a specialized meaning that takes them at least a page to explain, and so my attempt to summarize them in a paragraph will definitely miss out on important nuance. As well, after attempting to explain them normally, I'll try to view them through the lens of updating and feedback.

1. Commitment to Connection: remain in connection with the other despite resistance and impulses to break it, while not forcing yourself to stay when you genuinely want to separate or move away from the other. Reveal yourself to the other, and be willing to fully receive their expression before responding. This generates the high bandwidth information channel that can explore more broadly, while still allowing feedback; if you reveal an intense emotion, I let it land and then share my authentic reaction, allowing you to see what actually happens when you reveal that emotion, and allowing me to see what actually happens when I let that emotion land.
2. Owning Experience: Orient towards your impressions and emotions and stories as being *yours*, instead of about the external world. "I feel alone" instead of "you betrayed me." It also involves acknowledging difficult emotions, both to yourself and to others. The primary thing this does is avoid battles over "which interpretation is canonical," replacing that with easier information flow about how different people are experiencing things; it also is a critical part of updating about what's going on with yourself.

3. Trusting Experience: Rather than limiting oneself to emotions and reactions that seem appropriate or justifiable or ‘rational’, be with whatever is actually present in the moment. This gives you a feedback loop of what it’s like to follow your attention, instead of your story of where your attention should be, and lets you update that story. It also helps draw out things that are poorly understood, letting the group discover new territory instead of limiting them to territory that they’ve all been to before. It also allows for all the recursion that normal human attention can access, as well as another layer, of attending to what it’s like to be attending to the Circle when it’s attending to you.
4. Staying with the Level of Sensation: An echo of Commitment to Connection, this is about not losing touch with the sensory experience of being in your body (including embodied emotions) while speaking; this keeps things ‘alive’ and maintains the feedback loop between your embodied sense of things and your conscious attention. It has some similarities to Gendlin’s Focusing. Among other things, it lets you notice when you’re boring yourself.
5. Being with the Other in Their World: This one is harder to describe, and has more details than the others, but a short summary is “be curious about the other person, and be open to them working very differently than you think they work; be with them as they reveal themselves, instead of poking at them under a microscope.” This further develops the information channel, in part by helping it feel fair, and in part by allowing for you to be more surprised than you thought you would be.

Having said all that, I want to note that I might be underselling Commitment to Connection. The story I'm telling here is "Circling is powered in part by a methodological commitment to openness," and noting that science and rationality are powered similarly, but another story you could tell is "Circling is powered in part by a commitment to connection." That is, a scientist might say "yes, it's hard to learn that you're wrong, but it's worth it" and analogously a Circler might say "yes, it's hard to look at difficult things, but it's worth it," but furthermore a Circler might say "yes, it's hard to look at difficult things, but we're in this together."

Reflection as Secret Sauce

It’s one thing to have a feedback loop that builds techne, but I think Circling goes further. I think it taps into the power of reflection that creates a Lens That Sees Its Flaws. Humans can Circle, and humans can understand Circling; they can Circle about Circling. (They can also write blog posts about Circling, but that one’s a bit harder.) There’s also a benefit to meditating together, as I will have an easier time seeing my blind spots when they’re pointed out to me by other members of a Circle than when I go roaming through my mind by myself. Circling seems to be a way to widen your own lens, and see more of yourself, cultivating those parts to be more deliberate and reflective instead of remaining hidden and unknown.

Don't Double-Crux With Suicide Rock

Honest rational agents should never agree to disagree.

This idea is formalized in Aumann's agreement theorem and its various extensions (we can't foresee to disagree, uncommon priors require origin disputes, complexity bounds, *&c.*), but even without the sophisticated mathematics, a basic intuition should be clear: there's only one reality. Beliefs are for mapping reality, so if we're asking the same question and we're doing everything right, we should get the same answer. Crucially, even if we haven't seen the same evidence, the very fact that you believe something *is itself evidence* that I should take into account—and you should think the same way about my beliefs.

In "The Coin Guessing Game", Hal Finney gives a toy model illustrating what the process of convergence looks like in the context of a simple game about inferring the result of a coinflip. A coin is flipped, and two players get a "hint" about the result (Heads or Tails) along with an associated hint "quality" uniformly distributed between 0 and 1. Hints of quality 1 always match the actual result; hints of quality 0 are useless and might as well be another coinflip. Several "rounds" commence where players simultaneously reveal their current guess of the coinflip, incorporating both their own hint and its quality, and what they can infer about the other player's hint quality from their behavior in previous rounds. Eventually, agreement is reached. The process is somewhat alien from a human perspective (when's the last time you and an interlocutor switched sides in a debate *multiple times* before eventually agreeing?!), but not completely so: if someone whose rationality you trusted seemed visibly unmoved by your strongest arguments, you would infer that they had strong evidence or counterarguments of their own, even if there was some reason they couldn't tell you what they knew.

Honest rational agents should never agree to disagree.

In "Disagree With Suicide Rock", Robin Hanson discusses a scenario where disagreement seems clearly justified: if you encounter a rock with words painted on it claiming that you, personally, should commit suicide according to your own values, you should feel comfortable disagreeing with the words on the rock without fear of being in violation of the Aumann theorem. The rock is probably just a rock. The words are information from whoever painted them, and maybe that person *did* somehow know something about whether future observers of the rock should commit suicide, but the rock itself doesn't implement the dynamic of responding to new evidence.

In particular, if you find yourself playing Finney's coin guessing game against a rock with the letter "H" painted on it, you should just go with your own hint: it would be incorrect to reason, "Wow, the rock is still saying Heads, even after observing my belief in several previous rounds; its hint quality must have been *very* high."

Honest rational agents should never agree to disagree.

Human so-called "rationalists" who are aware of this may implicitly or explicitly seek agreement with their peers. If someone whose rationality you trusted seemed visibly unmoved by your strongest arguments, you might think, "Hm, we still don't agree; I should update towards their position …"

But another possibility is that your trust has been misplaced. Humans suffering from "algorithmic bad faith" are on a continuum with Suicide Rock. What matters is the counterfactual dependence of their beliefs on states of the world, not whether they know all the right keywords ("crux" and "charitable" seem to be popular these days), nor whether they can *perform the behavior* of "making arguments"—and *definitely not* their subjective conscious verbal narratives.

And if the so-called "rationalists" around you suffer from *correlated* algorithmic bad faith—if you find yourself living in a world of painted rocks—then it may come to pass that protecting the sanctity of your map requires you to master the technique of lonely dissent.

Criticism as Entertainment

ETA: This post relies on video embeds that didn't carry over when LW imported it. For the full experience, see here.

Media Reviews

There is a popular genre of video that consist of shitting on other people's work without any generative content. Let me provide some examples.

First, Cinema Sins. This is the first video I selected when looking for a movie I'd seen with a Cinema Sins I hadn't (i.e. it's not random, but it wasn't selected for being particularly good or bad).

The first ten sins are:

1. Use of a consistent brand for props in the movie they'd have to have anyway, unobtrusively enough that I never noticed until Cinema Sins pointed it out.
2. A character being mildly unreasonable to provoke exposition.
3. The logo
4. Exposition that wasn't perfectly justified in-story
5. A convenience about what was shown on screen
6. A font choice (from an entity that in-universe would plausibly make bad font choices)
7. An omission that will nag at you if you think about it long enough or expect the MCU to be a completely different thing, with some information about why it happened.
8. In-character choices that would be concerning in the real world and I would argue are treated appropriately by the movie, although reasonable people could disagree
9. Error by character that was *extremely obviously* intentional on the part of the film makers. There is no reasonable disagreement on this point.
10. An error perfectly in keeping with what we know about the character.

Of those, three to four could even plausibly be called sins of the movie- and if those bother you, maybe the MCU is not for you. The rest are deliberate choices by filmmakers to have characters do imperfect things. Everyone gets to draw their own line on characters being dumb- mine is after this movie but before 90s sitcoms running on miscommunication- but that's irrelevant to this post because Cinema Sins is not helping you determine where a particular movie is relative to your line. Every video makes the movie sound exactly as bad as the last, regardless of the quality of the underlying movie. It's like they analyze the dialogue sentence by sentence and look to see if there's anything that could be criticized about it.

Pitch Meeting is roughly as useful, but instead of reacting to sentences, it's reading the plot summary in a sarcastic tone of voice.

Pitch Meeting is at least bringing up actual problems with Game of Thrones season 8. But I dare you to tell if early Game of Thrones was better or worse than season 8, based on the Pitch Meeting.

I keep harping on “You can’t judge movie quality by the review”, but I don’t actually think that’s the biggest problem. Or rather, it’s a subset of the problem, which is you don’t learn anything from the review: not whether the reviewer considered the movie “good” or not, and not what could be changed to do make it better. Contrast with Zero Punctuation, a video game review series notorious for being criticism-as-entertainment, that nonetheless occasionally likes things, and at least once per episode displays a deep understanding of the problems of a game and what might be done to fix it.

Why Are You Talking About This?

It’s really, really easy to make something look bad, and the short-term rewards to doing so are high. You never risk looking stupid or having to issue a correction. It’s easier to make criticism funny. You get to feel superior. Not to mention the sheer joy in punishing bad things. But it’s corrosive. I’ve already covered (harped on) how useless shitting-on videos are for learning or improvement, but it goes deeper than that. Going in with those intentions changes how you watch the movie. It makes flaws more salient and good parts less so. You become literally less able to enjoy or learn from the original work.

Maybe this isn’t universal, but for me there is definitely a trade off between “groking the author’s concepts” and “interrogating the author’s concepts and evidence”. Groking is a good word here: it mostly means understand, but includes playing with the idea and applying it what I know. That’s very difficult to do while simultaneously looking for flaws.

Should it be, though? Generating testable hypotheses should lead to greater understanding and trust or less trust, depending on the correctness of the book. So at least one of my investigation or groking procedures are wrong.

The Alignment-Competence Trade-Off, Part 1: Coalition Size and Signaling Costs

This is part 1 of a series of posts I initially planned to organize as a massive post last summer on principal-agent problems. As that task quickly became overwhelming, I decided to break it down into smaller posts that ensure I cover each of the cases and mechanisms that I intended to.

Overall, I think the trade-off between the alignment of agents and the competence of agents can explain a lot of problems to which people often think there are simple answers. The less capable an agent is (whether the agent is a person, a bureaucracy, or an algorithm) the easier it is for a principal to assess the agent, and ensure the agent is working toward the principal's goals. As agents become more competent, they become both more capable of actually accomplishing the principal's goals and of merely appearing to accomplish the principal's goals while pursuing their own. In debating policy changes, I often find one sided arguments that neglect this trade-off, and in general I think efforts to improve policies or the bureaucratic structures of companies, non-profits, and governments should be informed by it.

Part 1:

Virtue signaling and moralistic anger are both forces that have been useful for holding people accountable, and powerful mechanisms of cultural evolution: spreading some norms more successfully than others, and resulting in many societies holding similar norms.

However, the larger a group becomes, the less members of the group know on average about other individual member's behavior or the consequences of it: making it harder to evaluate complex actions. This in turn gives an advantage to more clear forms of signaling that are more inefficient and costly than those that could be sustainable in smaller groups.

Examples:

- While it would be efficient for a politician to accept money from competing special interest groups and to keep their behavior consistent with their constituents regardless, it is simpler for politicians to convince allies they aren't corrupt by not accepting money from political opponents at all.
- With more complex tax codes, governments can implement more economically efficient pigouvian taxes which increase economic growth by concentrating more and more of the tax burden on actions which produce negative externalities for others. However, the power to assess and tax negative externalities gives those that influence tax code the opportunity to shape tax code to their own advantage at the expense of others.
- While a police officer could accept bribes and enforce the law anyway, unless you have a lot of information on the officer, you wouldn't be convinced that there weren't cases the officer was looking the other way. Likewise, people probably don't trust the objectivity of police departments with the power of civil

asset forfeiture even if such power can be used to reduce the tax burden of police and to create stronger deterrent effects on crime.
- More pacifistic states are more credible in not holding hostile expansionist intentions, than defensive states, who are in turn more credible than states that take offensive or pre-emptive action.
- It is simpler for someone to be vegan than to try explaining a series of edge cases about animal welfare/consciousness or a strategy of eating meat when it doesn't increase demand for meat. It would also look pretty suspicious for vegans to try selling wasted meat, even though doing so would undercut meat producers.
- It may be more efficient for you to work from home or to shift your hours on fairly autonomous work to match times you are more productive, but generally employers require explanations, and seek to avoid giving their employees room to slack off.
- Nepotistic hiring enables employers to ensure the alignment of employees, via additional information on prospective employees and leverage on their social capital. More meritocratic hiring is more ideal for society, however, the more candidates there are to assess the harder it becomes to investigate the merit of each. Accordingly, the larger a competing population of candidates is, the more their education will become focused on winning signaling competitions, and the less it will become focused on gaining skills that are difficult to demonstrate quickly.

In summary there are a lot of actions that are more directly efficient and selfishly beneficial for those that do them, but because they are not credible signals of good intent/are excuses that the corrupt would use, the options are not sustainable in larger societies. Small groups where people know each other well on the other hand can allow weirder norms to be sustainable without corruption due to their increased ability to vet each other. This may also explain why smaller groups in history often had more sustainable norms of exploiting defectors or outsiders which wouldn't be sustainable in larger societies since you can't tell if someone is robbing a thief or an innocent person. Reducing attempts at exploitation between small competing groups of insiders is likewise probably a good thing for scaling up societies.

In general, these signaling costs come from scenarios where people's interests may not align, and the costs are paid to demonstrate alignment. Without efficient mechanisms to assess and vet each other, as groups scale they lose trust, and more costly signaling becomes required to sustain cooperation.

Studying Early Stage Science: Research Program Introduction

This is a linkpost for https://medium.com/@LeverageResearch/early-stage-science-research-program-introduction-cb52b7c5d6f1

(Leverage just posted the below introduction into their current research program to Medium, and I got permission to crosspost it here. I think it's quite a good read, and I more broadly think that the history of science is one of the best ways to study the art of rationality)

Introduction

Scientific progress is responsible for some of the most amazing developments in human history. It has enabled us to cure diseases, increase agricultural yields, and travel quickly and safely across the globe. As such, the opportunity to enable progress in science can be very valuable.

Having studied examples from the history of science, we at Leverage Research believe it is possible to describe how early discoveries led to the creation of impressive and sophisticated scientific disciplines. By understanding the methodologies used by the researchers at key points in the past, as well as the social and institutional contexts that enabled them to make progress, we believe it may be possible to help modern researchers make progress in new or stagnating fields.

In this paper, we outline a research program designed to investigate this hypothesis. We describe three important examples from the history of science, identify a phenomenon worth studying, state a specific formulation of our hypothesis, and describe our research methodology[1]. With this paper and the work that follows it, we hope to set the stage for the study of early stage science.

You can read more about early stage science on our website.

Three Historical Cases

Galvani, Volta, and the Invention of the Battery

In 1780, an Italian physician and biologist named Luigi Galvani announced the discovery of a new electrical phenomenon, which he called “animal electricity.”

For some time, Galvani had been performing experiments using dissected frogs to investigate how electricity interacted with animal bodies. In his most famous experiment, he found that he could make dissected frog legs twitch by hanging them on iron hooks and probing them with a piece of metal. He concluded on the basis of his investigations that an electrical charge was being generated by the frog leg itself, and that this “animal electricity” was generated by an electrical fluid in the frogs (Cajavilca, Varon, and Sternbach 2009, 160).

Figure 1: Galvani experimenting on frogs

Galvani's research attracted others to investigate animal electricity. In particular, it drew the attention of an Italian physicist and chemist by the name of Alessandro Volta.

Volta began to suspect that electricity was being generated from a source other than the frog itself. Volta's hypothesis was that the electrical charge was instead created by Galvani's use of different metals to mount and probe the frog leg, and further that certain metals were naturally disposed to pass an electric charge given the presence of a conductor.

To test his hypothesis, Volta needed some way to detect the presence of an electric charge. In the days before the invention of precise instruments for measuring electricity, this was no small challenge. Others had used the creation of visible sparks or the physical sensation of getting an electric shock as a way of detecting electricity, but these effects tended to require a strong current and Volta suspected that the current involved in the frog leg experiment was relatively small.

He needed a more sensitive instrument. So, he used his tongue.

He reasoned that if a frog leg would conduct electricity, his tongue probably would too. And, Volta knew from previous experiments that an electric charge applied to the tongue could

create a bitter sensation. This let him conduct an experiment: he touched either side of his tongue with combinations of metals and used the presence or absence of the bitter sensation to detect an electrical charge. He found that some combinations of metals did cause the bitter sensation, leading him to conclude that it was in fact the metals that passed a current through his tongue, rather than his tongue generating its own “animal electricity” (Cajavilca, Varon, and Sternbach 2009, 162; Shock and Awe: The Story of Electricity 2011).

Subsequent investigations led Volta to make further breakthroughs in understanding the ability of metals to transmit an electric charge. He eventually discovered that he could generate a strong and consistent electrical charge by stacking zinc and copper on top of one another in an alternating pattern with brine-soaked cloth or cardboard in between.

Volta’s invention was dubbed the Voltaic pile, and was the precursor to the modern battery.

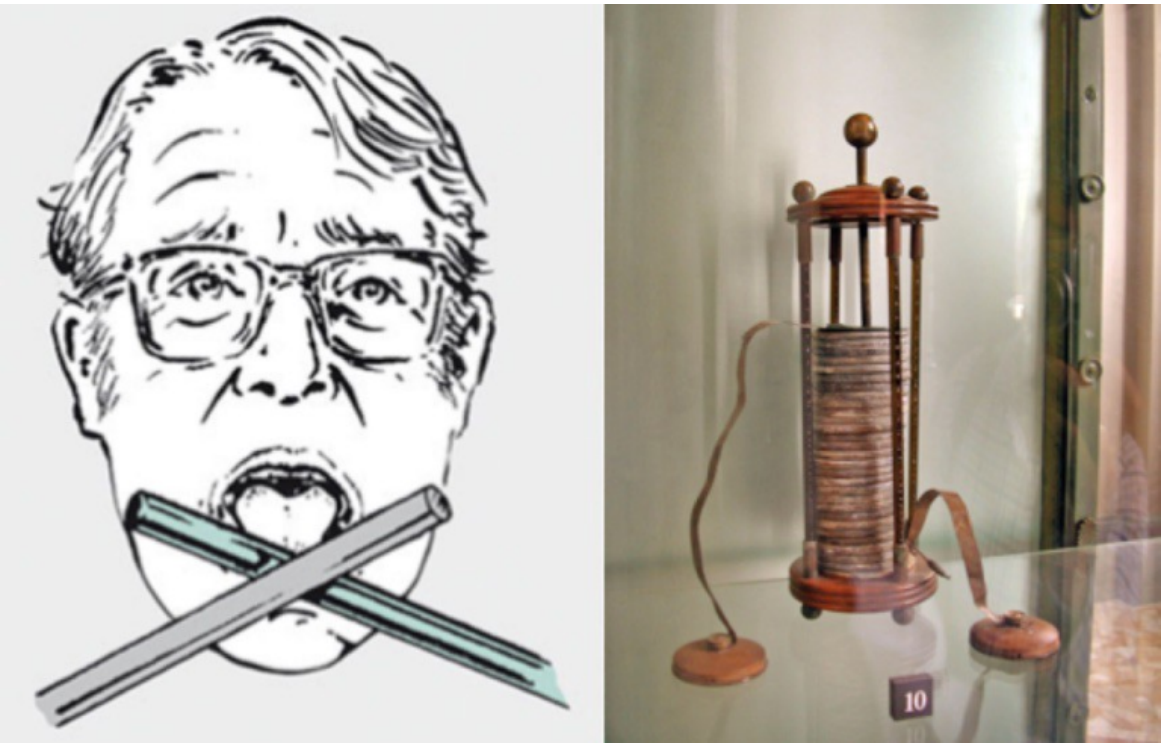


Figure 2: Volta’s experiment and the Voltaic Pile

The debate between Galvani and Volta over animal electricity and related topics would result not only in the Voltaic pile, but also in the early development of the fields of electrophysiology, electromagnetism, and electrochemistry (Cajavilca, Varon, and Sternbach 2009, 159).

While the two had deep theoretical disagreements, Galvani’s creative experiments identified a phenomenon — frogs twitching when touching two pieces of metal — which Volta was able to study with a unique method of his own, using his tongue instead of a frog. Galvani’s theory of animal electricity was wrong, but his work enabled further research that led to the creation of the voltaic pile and the foundations of our modern understanding of electricity.

Volta and Galvani both had to develop many of their own theories in the absence of an accepted paradigm for electricity, and both had to design their own experimental tools and methods. The trajectory of their respective research programs, and of their research efforts combined, would have been hard to predict in advance. They tried anyway, and their work helped turn the study of electrical phenomenon into a fully developed scientific discipline.

Galileo and the Development of Telescopic Astronomy

The history of science includes a number of cases of scientific progress that share some of the notable features of Galvani and Volta's research. Another example is the story of Galileo and the early development of telescopic astronomy.

Galileo Galilei pioneered the use of the telescope to observe celestial objects. He observed Jupiter's moons, the phases of Venus, Saturn, the existence of sunspots and more. Initially, though, Galileo had trouble convincing others of the reliability and usefulness of the telescope. When Galileo visited Bologna in 1610 to demonstrate his telescope:

> *In the presence of a number of learned men, Galileo showed his telescope and let others observe earthly and celestial things through it. They agreed that for earthly objects the instruments performed as promised but that in the heavens it was not reliable. Although Galileo's notes show that on the first night two and on the second night all four of the satellites were visible, none of the gentlemen present were able to see satellites around Jupiter. (Van Helden 1994, 11)*

The learned men were skeptical of Galileo's device for understandable reasons. The optical principles behind the device were not well understood at the time and some of Galileo's claims about the heavens contravened established wisdom on the topic. Furthermore, Galileo's telescope was not a very powerful or easy-to-use instrument. His telescope was capable of no more than 20x magnification and a field of view of around 15 feet, whereas one can purchase a modern amateur telescope today for around 100 USD that is capable of 10 times greater magnification (200x) and a field of view 34 times larger (516 feet) (Sky and Telescope 2017; Telescopic Watch 2019).

To illustrate the difficulties this posed, compare the image of Ely Cathedral viewed with a modern telescope at the same zoom as Galileo's telescope (20x) with the field of view available to modern telescopes (Figure 3, top picture), with the image of the same object at the same distance with a replica of one of Galileo's telescopes (Figure 3, bottom picture).

Figure 3: Top picture — modern telescope at 20x zoom; Bottom picture — replica of Galileo's telescope, also at 20x zoom (Astronomy and Nature TV 2010)

The small circle in the middle of the bottom picture is the field of view available through the replica telescope. The light colored area shows the visible portion of the cathedral.

In addition to the small field of view, Galileo lacked modern equipment for mounting and stabilizing his telescope. This meant that even if Galileo could find the relevant astronomical

object in the small field of view, ensuring that the telescope remained fixed on that object for repeated observations was a significant challenge.

Finally, even under optimal circumstances, it is not uncommon for a user of Galileo's telescope to fail to see what they are meant to see. The instrument takes a certain amount of getting used to as the eye and brain become accustomed to interpreting the image and not everyone has sufficiently good eyesight to use the device properly (Van Helden 1994, 11). As one historian of science notes:

> *On several occasions I have taken a group of my students to look for Jupiter's satellites through a replica of one of Galileo's telescopes — students who were convinced the moons were really there — and the results were always mixed. Some saw all that were visible, some saw one or two, and some saw none at all. No matter how public the occasion, the actual observing remains an individual and private act. (Van Helden 1994, 12)*

Indeed, the private nature of telescopic observations created difficulties in settling scientific disputes. For example, in 1643, Antonius Rheita published a book announcing the discovery of new satellites around Jupiter and Saturn. While others reported not seeing the satellites, Rheita claimed the satellites could only be seen with the new, more powerful telescope he had invented. All telescopes at the time were hand-made and thus differed in clarity and magnification, and so scientists could not simply use their own instruments to confirm or disconfirm each others' claims as their failure to replicate results could be the consequence of inferior equipment. Scientists eventually reached agreement on this question in 1647, when Hevelius (who claimed to have built a still more powerful telescope), was able to convince scientists that the supposed satellites were actually fixed stars behind the planets (Van Helden 1994, 8–20).

The telescope was initially an unreliable, poorly understood tool. It did not produce directly shareable or replicable results and in practice often created scientific disputes that were difficult to resolve. Yet despite the challenges it posed, the telescope did sometimes allow early telescopic astronomers to make observations that were more detailed than those available to the naked eye. This modest improvement was enough to attract a small number of early scientists who began working with the device. Over time, they learned more about the optical principles involved, improved at grinding lenses for new telescopes, and developed more effective ways of communicating their findings to other astronomers and the public.

Black, Lavoisier, and the Chemical Revolution

In 1756, Joseph Black published his Experiments Upon Magnesia Alba in which he described experiments he performed on an unusual substance which he called "fixed air." Black originally discovered the substance[2] through experiments on magnesia alba — now called magnesium carbonate — and chalk (calcium carbonate). Black observed that when magnesia alba was heated or combined with an acid it began to bubble and left behind a residue. Black was able to use an analytical balance that he invented to precisely weigh the residue and note that it lost a noticeable amount of weight. Black hypothesized that the bubbling and weight loss was caused by the liberation of a gas that had been "fixed" in the magnesia alba (hence the name, fixed air).

Black was curious about the properties of this fixed air and so began to devise ways to experiment with the substance.

He noticed that it had an unusual effect on fire, noting that:

> *I mixed together some chalk and vitriolic acid. . . The strong effervescence produced an air or vapour, which, flowing out at the top of the glass, extinguished a candle that stood*

> *close to it; and a piece of burning paper immersed in it, was put out as effectually as if it had been dipped in water. (West 2014, L1059)*

This indicated that fixed air was not the same as atmospheric air. There were other unusual properties as well. He investigated the effect of fixed air on animals and found that it was remarkably toxic when inhaled — he noted that “sparrows died in it in ten or eleven seconds” (Robinson 1803).

The toxicity of fixed air when inhaled led Black to suspect that fixed air might be the air expelled as part of the respiration process itself. He designed several experiments to test this hypothesis. In one he blew bubbles into a solution of limewater (calcium hydroxide) and noted that a precipitate of chalk was leftover indicating that he had succeeded in fixing the air back into the chalk. He repeated this experiment on a larger scale by placing limewater soaked in rags in an air duct in the ceiling of a church and observing a chalk residue as the lime soaked up the fixed air from the congregation’s respiration (West 2014, L1059).

He made several other discoveries about the substance and explained various observed phenomena. For instance, there was an observed unusual effect at the Grotto del Cano in Italy, where animals that visited the Grotto died, but humans could visit unharmed. Black explained this phenomenon by proposing that fixed air was heavier than atmospheric air and thus sank closer to the ground, poisoning animals but not humans. He also discovered that fixed air was emitted during the fermentation process (West 2014, L1059).

We now know “fixed air” as carbon dioxide, and Black’s work isolating and describing the properties of carbon dioxide represented the first demonstration that gases can be weighable constituents of solid bodies, the first demonstration that gases are unique chemical substances and not atmospheric air in different states of purity, and the first demonstration that respiration involves the transformation of gases. In the 18 years after the publication of Black’s work all the respiratory gases were isolated and characterized with Henry Cavendish discovering hydrogen in 1766, Daniel Rutherford isolating nitrogen in 1772, and Joseph Priestly isolating oxygen in 1774 (West 2014, L1059).

Joseph Black’s work in isolating and describing carbon dioxide contributed significantly to scientific progress in chemistry. In much the same fashion as Galvani and Volta, Black discovered a new phenomenon and developed unusual and creative methods to study the phenomenon. In particular, his insight that carbon dioxide was connected to respiration allowed for rapid progress in isolating the other gases involved in respiration. As other scientists investigated new gases they attempted to explain them in terms of ultimately incorrect, yet sometimes useful theories. Indeed, even Black’s theory that the air was “fixed” in the solids is somewhat different than the modern understanding of the phenomenon. In modern nomenclature, Magnesia Alba is MgCO3, and by heating it MgCO3 becomes MgO + CO2. We might say that carbon dioxide can be liberated from Magnesia Alba but in modern terms, we probably wouldn’t describe carbon dioxide as being “fixed” in the Magnesia Alba. While Black’s theory of carbon dioxide may not match our modern understanding, he nevertheless contributed significantly to scientific progress by isolating and cleverly studying a new phenomenon.

As additional gases were isolated and described, there was considerable disagreement about what the gasses were and scientists advanced a number of competing theories to explain their properties. For example, Rutherford and Priestly’s discoveries were both explained in terms of the then popular phlogiston theory of combustion which posited that an element called phlogiston was released from materials as they combusted and combustion would continue until either all the phlogiston had been released or the air was saturated with phlogiston such that it could not contain more. On this theory nitrogen was labeled “phlogisticated air” and oxygen was labeled “dephlogisticated air.”[3]

The debate over the nature of these newly discovered airs culminated in the work of Antoine-Laurent de Lavoisier and what is often called the chemical revolution. Between 1775 and

1789, Lavoisier is credited with discovering the law of conservation of mass and a new theory of combustion, which explained combustion and acidic corrosion in terms of oxygen[4] and eventually replaced the phlogiston theory. Lavoisier's approach to chemistry built on Black's approach through its focus on weight, but utilized much more sophisticated and elaborate equipment to investigate the properties of the newly discovered airs (West 2014, L1060).

Identifying a Phenomenon

The research efforts of Volta and Galvani, Galileo, and Black appear to have a number of attributes in common.

They feature a relative absence of established theories and well-understood instruments in the area of investigation, the appearance of strange or unexplained phenomena, and lack of theoretical and practical consensus among researchers. Progress seems to occur despite (and sometimes enabled by) flawed theories, individual researchers use imprecise measurement tools that are frequently new and difficult to share, and there exists a bi-directional cycle of improvement between increasingly sophisticated theories and increasingly precise measurement tools.

For example, consider Black's study of fixed air. He began with an unexplained phenomenon: certain substances losing weight when heated. His approach to studying that phenomenon involved a wide range of methods, such as blowing bubbles into limewater and leaving limewater-soaked rags in the air duct of a church. This led to a larger study of new gases, where there were many theoretical disagreements between researchers, and where flawed theories (e.g., the theories of phlogisticated and dephlogisticated air) seemed to aid the process of discovery. Finally, improvements in measuring the weight of substances allowed for better theories and refinements to the researchers' instruments.

These examples gesture at the potential existence of a recognizable cluster of discovery-related attributes ("Attribute Cluster 1") that plausibly play an important role in scientific progress. This is striking, because this is different from another attribute cluster ("Attribute Cluster 2") that is more familiar and more commonly referred to, that appears throughout the history of science. That second cluster includes large groups of scientists working together, a foundation of largely accepted theory, precise and well-understood instruments, researcher consensus on the quality of these instruments, and many small discoveries that tend to build iteratively on one-another and cohere with previous theory.

This raises a number of questions: Are the attributes in Attribute Cluster 1 actually present in cases described above? Do those attributes appear in a special set of cases in the history of science? More generally, is there a natural cluster of discovery-related attributes in the conceptual vicinity of Attribute Cluster 1 that appear in important cases of scientific discovery, like those described above? What can we learn from investigating the cases that are the most natural candidates for exemplifying attributes from Attribute Cluster 1? Is the distinction between Attribute Cluster 1 and Attribute Cluster 2 useful in the context of describing the progress of science?

The pattern suggested by the cases described above, and those like them, is noteworthy. If there is in fact a recognizable cluster of discovery-related attributes that plausibly play a role in scientific progress other than Attribute Cluster 2, this could be important for understanding when and how new fields make scientific progress. We believe this possibility is rendered at least somewhat plausible by the cases above and deserves further study.

Hypothesis

The cases of Volta and Galvani, Galileo, and Black illustrate a hypothesized pattern in the development of fields of science which we call “early stage science.”

We hypothesize that:

1. Some scientific fields develop from initial investigations in nascent fields to highly functional knowledge acquisition programs.
2. The histories of the development of these highly functional knowledge acquisition programs are characterized by a similar, describable pattern.
3. As part of this pattern, the relevant scientific fields have different attributes at different points in their development. More specifically, earlier in the development of the relevant scientific fields, the fields have an attribute cluster in the conceptual vicinity of Attribute Cluster 1, as suggested by numerous examples. Some of those fields later have an attribute cluster in the conceptual vicinity of Attribute Cluster 2. The fact of these attribute clusters implies the existence of phases of scientific development.
4. These phases of scientific development arise from the facts about how people figure things out about the world. This includes how inference, experimentation, observation, theory development, tool development, and other aspects of how people can seek to understand the world around them lead researchers to be able to improve their understanding of a given phenomenon.
5. In accordance with this, researchers need to use different tools and practices depending on their starting state of knowledge. We expect to find that there is a coherent logic to the meta-practices that lead to the development and advancement of research programs under different starting conditions.
6. More specifically, a model of early stage scientific practice will explain how researchers overcome the difficulties that arise when attempting to gain knowledge when dealing with substantially unknown phenomena, poor tools, and so forth. When stated, this model will be both prima facie plausible and verifiable by checking against the research activities that led to important discoveries in practice.

It is consistent with the above hypothesis that many of the attributes specifically identified in the historical cases above are not actually constitutive of early stage science, so long as conceptually similar attributes can be identified which are.

Methodology

The current methodology of our investigation primarily focuses on analyzing historical case studies of scientific discovery in new fields. In this section, we cover how we plan to select and analyze specific cases.

Selecting Cases

To select cases, we will attempt to identify functional modern and historical scientific research programs and identify the initial discoveries integral to their development. By starting with functional scientific research programs and working back, we are aiming to ensure that the cases of early stage research we study are success cases. The pattern that we hypothesize to exist is not meant to be a pattern that occurs in all early stage research, but rather effective early stage research. Since the effectiveness of early stage research can be difficult to assess, we believe the safest set of cases to examine are cases where the relevant field developed into a full-fledged functional research program.

To identify functional modern and historical scientific research programs, we will look for fields where there are large groups of researchers studying similar phenomena, using very similar methods, with a corpus of shared theory, and substantial predictive power. This is our initial hypothesis of signs that will enable us to identify a sufficient set of functional scientific research programs. If we encounter reasons to change our criteria, we will do so, especially if

doing so will help us to better identify successful early stage research cases. It is open to us, for instance, that we would be better served by looking at Kuhnian paradigms or broad scientific consensus, rather than looking for the attributes above.

Once we identify functional scientific research programs and identify the initial discoveries integral to their development, we will then analyze how researchers made those discoveries.

For example, consider modern astronomy. As part of doing their research, large groups of modern astronomers use highly advanced telescopes and highly advanced analytical methods to find, see, and study intergalactic objects at an incredible degree of precision. Astronomers have a shared body of theory, make the same observations with different telescopes, and are able to predict with high precision what they and others will observe when they look through telescopes at different points in the sky. These attributes indicate that modern astronomy is plausibly a highly functional scientific research program.

Having selected astronomy as a plausible success case for scientific development, we can then investigate the beginnings of the field to find plausibly important and formative discoveries. In this case, Galileo and others' work with early telescopes significantly changed how we both observe and think about objects in space. As a result, Galileo plausibly represents a case where successful early research methods might be visible and thus is fruitful to study.

Analyzing Cases

In analyzing the historical cases, we will be trying to build a coherent model of how researchers make scientific progress in the early stages of the development of scientific fields. We expect to approach this from many angles, looking at similarities and differences between cases, building causal models of the cases, and trying to state plausible general obstacles early stage researchers might encounter. By cross-checking the cases and the general model 11 against each other, we hope to improve our models of individual cases as well as our general model.

This may or may not lead to convergence on a single, intelligible general model that fits the cases and also is plausible abstractly. If it does, we will consider this evidence in favor of our hypothesis. If it does not, either because plausible general models do not fit the cases or because the general models contain attributes that are not intelligibly related to the logic of discovery in the early stages of a field, we will consider this evidence against our hypothesis.

If we reach an adequately plausible general model, we will then investigate whether it can be used to generate recommendations for present-day scientists seeking to make progress in fields operating under early stage research conditions. If a particular model relies heavily on factors local to the particular historical case, researcher, or era, it may not serve as a good template for recommendations for present-day scientists or a general theory of scientific methodology. On the other hand, if a model tends to feature potentially generalizable methodological elements with mechanistic relations (e.g., a particular model for how instrument development works at different levels of theoretical uncertainty), we can see whether the model can generate recommendations for current researchers.

Conclusion

Following the methodology above, we believe we will be able to select and analyze cases of discovery that have led to the development of highly functional scientific research programs. We expect our analysis to help confirm or disconfirm the existence of early stage science as a unique phase of science with a unique and understandable methodology. It is our hope that investigating this hypothesis will shed light on what methods should be used to make scientific progress in new or underdeveloped fields, and thereby help push science forward.

Footnotes

1. A later paper will situate our investigation in the surrounding literature, including describing our relation to Kuhn, Popper, Lakatos, historicism, methodologism, natural epistemology, and complementary science.
2. Black was likely not the original discoverer of the substance. It had been briefly described around 100 years prior by Jan Baptist van Helmont who called it gas sylvestre although Black was likely the first to describe its properties in detail.
3. Rutherford's initial name for nitrogen was "noxious air." Labeling it phlogisticated air is often attributed to Priestly.
4. Indeed, the name "oxygen" comes from the greek roots (oxys) meaning "acid" and (-gen¯es) meaning "producer."

This paper was originally posted on our website.

References

- Bartlett, Richard J. 2019. "Galileo's Telescope — The What, When and How." Telescopic Watch. Last modified May 16, 2019. https://telescopicwatch.com/galileo-telescope/.
- Cajavilca, Christian, Joseph Varon, and George L. Sternbach. 2009. "Luigi Galvani and the Foundations of Electrophysiology." Resuscitation 80 (2): 159–62. doi:10.1016/j.resuscitation.2008.09.020.
- Dalby, Robert. 2010. "Looking through Galileo's Telescope — Practical Comparison." Astronomy and Nature TV. Last modified September 27, 2019. 12 https://www.youtube.com/watch?v=nzXnnwxJmSg.
- Quinn, Jim. 2017. "Stargazing with Early Astronomer Galileo Galilei." Sky Telescope. Last modified May 9, 2019. https://www.skyandtelescope.com/astronomy-resources/stargazing-with-galileo/.
- Robison, John. 1803. Lectures on the elements of chemistry by the late Joseph Black. Edinburgh: W. Creech.
- "Spark." 2011. Shock and Awe: The Story of Electricity. BBC Four. Van Helden, Albert. 1994. "Telescopes and Authority from Galileo to Cassini." Osiris 9: 8–29. doi:10.1086/368727.
- West, John B. 2014. "Joseph Black, Carbon Dioxide, Latent Heat, and the Beginnings of the Discovery of the Respiratory Gases." American Journal of Physiology-Lunch Cellular and Molecular Physiology 306 (12): L1057–63.

Homeostasis and “Root Causes” in Aging

Let’s start with a stylized fact: almost every cell type in the human body is removed and replaced on a regular basis. The frequency of this turnover ranges from a few days (for many immune cells and cells in the gastrointestinal lining) to ten years (for fat, heart, and skeleton cells). Only a handful of tissues are believed to be non-renewing in humans - e.g. eggs, neurons, and the lens of the eye (and even out of those, neurons are debatable).

This means that the number of cells of any given type is determined by “homeostatic equilibrium” - the balance of cell removal and replacement. If an ulcer destroys a bunch of cells in your stomach lining, they’ll be replaced over a few days, and the number of stomach cells will return to roughly the same equilibrium level as before. If a healthy person receives a bunch of extra red blood cells in a transfusion, they’ll be broken down over a few months, and the number of blood cells will return to roughly the same equilibrium level as before.

As organisms age, we see a change in the homeostatic equilibrium level of many different cell types (and other parameters, like hormone and cytokine levels). In particular, a wide variety of symptoms of aging involve “depletion” (i.e. lower observed counts) of various cell types.

However, human aging happens on a very slow timescale, i.e. decades. Most cell counts equilibrate much faster - for instance, immune cell counts equilibrate on a scale of days to weeks. So, suppose we see a decrease in the count of certain immune cells with age - e.g. naive T cells. Could it be that naive T cells just wear out and die off with age? No - T cells are replaced every few weeks, so a change on a timescale of decades cannot be due to the cells themselves dying off. If the count of naive T cells falls on a timescale of decades, then either (a) the rate of new cell creation has decreased, or (b) the rate of old cell removal has increased (or both). Either of those would require some “upstream” change to cause the rate change.

More generally: in order for cell counts, or chemical concentrations, or any other physiological parameter to decrease/increase with age, at least one of the following must be true:

- the timescale of turnover is on the order of decades (or longer)
- rate of removal increases/decreases
- rate of creation decreases/increases

If none of these is true, then any change is temporary - the cell count/concentration/whatever will return to the same level as before, determined by the removal and creation rates.

Of those three possibilities, notice that the second two - increase/decrease in production/removal rate - both imply some other upstream cause. Something else must have caused the rate change. Sooner or later, that chain of cause-and-effect needs to bottom out, and it can only bottom out in something which equilibrates on a timescale of decades or longer. (Feedback loops are possible, but if all the components equilibrate on a fast timescale then so will the loop.) *Something* somewhere in the system is out-of-equilibrium on a timescale of decades. We’ll call

that thing (or things) a “root cause” of aging. It’s something which is not replaced on a timescale faster than decades, and it either accumulates or decumulates with age.

Now, the main criteria: **a *root* cause of aging *cannot* be a higher or lower value of any parameter subject to homeostasis on a faster timescale than aging itself**. Examples:

- Most cell types turn over on timescales of days to months. “Depletion” of any of these cell types cannot be a root cause of aging; either their production rate has decreased or their removal rate has increased.
- DNA damage (as opposed to mutation) is normally repaired on a timescale of hours - sometimes much faster, depending on type. “Accumulation” of DNA damage cannot be a root cause of aging; either the rate of new damage has increased or the repair rate has decreased.
- DNA mutations cannot be repaired; from a cell’s perspective, the original information is lost. So mutations *can* accumulate in a non-equilibrium fashion, and are a plausible root cause under the homeostasis argument.

Note that the homeostasis argument does *not* mean the factors ruled out above are not links in the causal chain. For instance, there’s quite a bit of evidence that DNA damage does increase with age, and that this has important physiological effects. However, there must be changes further up the causal chain - some other long-term change in the organism’s state leads to faster production or slower repair of DNA damage. Conversely, the homeostasis argument does not imply that “plausible root causes” *are* the true root causes - for instance, although DNA mutations could accumulate in principle, cells with certain *problematic* mutations are believed to be cleared out by the immune system - so the number of cells with these mutations is in equilibrium on a fast timescale, and cannot be a root cause of aging.

For any particular factor which changes with age, key questions are:

1. Is it subject to homeostasis?
2. If so, on what timescale does it turn over?
3. If it is subject to homeostasis on a timescale faster than aging, then what are the production and removal mechanisms, and what changes the production and removal rates with age?

These determine the applicability of the homeostasis argument. Typically, anything which can normally be fixed/replaced/undone by the body will be ruled out as a root cause of aging - the timescale of aging is very long compared to practically all other physiological processes. We then follow the causal chain upstream, in search of plausible root cause.

What we Know vs. How we Know it?

Two weeks ago I said:

> The other concept I'm playing with is that "what we know" is inextricable from "how we know it". This is dangerously close to logical positivism, which I disagree with my limited understanding of. And yet it's really improved my thinking when doing historical research.

I have some more clarify on what I meant now. Let's say you're considering my ex-roommate, person P, as a roommate, and ask me for information. I have a couple of options.

Scenario 1: I turn over chat logs and video recordings of my interactions with the P.

E.g., recordings of P playing music loudly and chat logs showing I'd asked them to stop.

Trust required: that the evidence is representative and not an elaborate deep fake.

Scenario 2: I report representative examples of my interactions with P.

E.g., "On these dates P played music really loudly even when I asked them to stop."

Trust required: that from scenario 1, plus that I'm not making up the examples.

Scenario 3: I report summaries of patterns with P

E.g., "P often played loud music, even when I asked them to stop"

Trust required: that from scenario 2, plus my ability to accurately infer and report patterns from data.

Scenario 4: I report what a third party told me

E.g. "Mark told me they played loud music a lot"

Trust required: that from scenario 3, plus my ability to evaluate other people's evidence

Scenario 5: I give a flat "yes good" or "no bad" answer.

E.g., "P was a bad roommate."

Trust required: that from scenario 3 and perhaps 4, plus that I have the same heuristics for roommate goodness that you do.

The earlier the scenario, the more you can draw your own conclusions and the less trust you need to have in me. Maybe you don't care about loud music, and a flat

yes/no would drive you away from a roommate that would be fine for you. Maybe I thought I was clear about asking for music to stop but my chat logs reveal I was merely hinting, and you are confident you'll be able to ask more directly. The more specifics I give you, the better an assessment you'll be able to make.

Here's what this looks like applied to recent reading:

Scenario 5: Rome fell in the 500s AD.

Even if I trust your judgement, I have no idea why you think this or what it means to you.

Scenario 4: In Rome: The Book, Bob Loblaw says Rome Fell in the 500s AD.

At least I can look up why Bob thinks this.

Scenario 3: Pottery says Rome fell between 300 and 500 AD.

Useful to experts who already know the power of pottery, but leaves newbies lost.

Scenario 2: Here are 20 dig sites in England. Those dated before 323 (via METHOD) contain pottery made in Greece (which we can identify by METHOD), those after 500 AD show cruder pottery made locally.

Great. Now my questions are "Can pottery evidence give that much precision?" and "Are you interpreting it correctly?"

Scenario 1: Please enjoy this pile of 3 million pottery shards.

Too far, too far.

In this particular example (from *The Fall of Rome*), 2-3 was the sweet spot. It allowed me to learn as much as possible with a minimum of trust. But there's definitely room in life for 4; you can't prove everything in every paper and sometimes it's more efficient to offload it.

I don't view 5 as acceptable for anything that's trying to claim to be evidenced based, or at least, any basis besides "Try this and see if it helps you." (which is a perfectly fine basis if it's cheap).

Does GPT-2 Understand Anything?

Some people have expressed that “GPT-2 doesn’t understand anything about language or reality. It’s just huge statistics.” In at least two senses, this is true.

First, GPT-2 has no sensory organs. So when it talks about how things look or sound or feel and gets it right, it is just because it read something similar on the web somewhere. The best understanding it could have is the kind of understanding one gets from reading, not from direct experiences. Nor does it have the kind of understanding that a person does when reading, where the words bring to mind memories of past direct experiences.

Second, GPT-2 has no qualia. This is related to the previous point, but distinct from it. One could imagine building a robotic body with cameras for eyes and microphones for ears that fed .png and .wav files to something like GPT-2 rather than .html files. Such a system would have what might be called experiences of the world. It would not, however, create an direct internal impression of redness or loudness, the ineffable conscious experience that accompanies sensation.

However, this is too high a bar to rule out understanding. Perhaps we should call the understanding that comes from direct personal experience “real understanding” and the kind that comes solely from reading with no connection to personal experience “abstract understanding.” Although I can’t “really understand” what it was like to fight in the Vietnam War (because I wasn’t there, man) I can still understand it in an abstract sense. With an abstract understanding, here are some things one can do:

- answer questions about it in one’s own words
- define it
- use it appropriately in a sentence
- provide details about it
- summarize it

Professional teachers distinguish between tests of knowledge (which can be handled by mere memorization) and tests of understanding, with the latter being more difficult and useful (see Bloom’s Taxonomy). Understanding requires connecting a new idea to ideas a student is already familiar with.

GPT-2 is able to pass many such tests of understanding. With an appropriate prompt (such as giving examples of what form the answer to a question should take) it is able to answer questions, define terms, use words appropriately in a sentence, provide details, and summarize.

This is understanding for most practical purposes. It shows that when GPT-2 uses a word, that word has the appropriate kinds of connections to other words. The word has been integrated into a large graph-like structure of relationships between what can reasonably be called concepts or ideas. When probabilities for the next token

have been generated, it has a certain propensity for using a particular word; but if that word is artificially blocked, other ways of saying the same thing also have been activated and will be used instead. It is reasonable to interpret this as having an “idea” of what it “wants” to “say” and at some point the quotation marks are no longer helpful, and we may as well dispense with them.

Here is an example. I input the following prompt into GPT-2 1.5B, with top-k=10 sampling:

"Indiana Jones ducked as he entered the cave to avoid being decapitated." In this sentence, the word "decapitated" means

Here are the first 10 results (truncated after the first sentence):
• "to be cut down" as well as "to be slain."
• "to chop off".
• "to cut off one of the branches of a tree."
• "The captain of the ship was killed in the cave."
• "to cut off, cut off by decapitation."
• "cut off".
• "cut off."
• to be "sliced off."
• "to be killed," which is the same thing as "to be killed by the sword."
• to fall from high altitude or to be cut down.
• "to have a head chopped off."

The system has a strong notion that “decapitated" means “to cut off” and “to kill” but is less likely to mention that the word has anything to do with a head. So its concept of “decapitation” appears to be approximately (but not completely) right. When prompted to write a sentence using the word “decapitate,” the sentences the system usually generates are consistent with this, often being used in a way consistent with killing, but only rarely mentioning heads. (This has all gotten rather grisly.)

However, one shouldn't take this too far. GPT-2 uses concepts in a very different way than a person does. In the paper “Evaluating Commonsense in Pre-trained Language Models,” the probability of generating each of a pair of superficially similar sentences is measured. If the system is correctly and consistently applying a concept, then one of the two sentences will have a high probability and the other a low probability of being generated. For example, given the four sentences

1. People need to use their air conditioner on a hot day.
2. People need to use their air conditioner on a lovely day.
3. People don’t need to use their air conditioner on a hot day.
4. People don’t need to use their air conditioner on a lovely day.

Sentences 1 and 4 should have higher probability than sentences 2 and 3. What they find is that GPT-2 does worse than chance on these kinds of problems. If a sentence is likely, a variation on the sentence with opposite meaning tends to have similar

likelihood. The same problem occurred with word vectors, like word2vec. “Black” is the opposite of “white,” but except in the one dimension they differ, nearly everything else about them is the same: you can buy a white or black crayon, you can paint a wall white or black, you can use white or black to describe a dog’s fur. Because of this, black and white are semantically close, and tend to get confused with each other.

The underlying reason for this issue appears to be that GPT-2 has only ever seen sentences that make sense, and is trying to generate sentences that are similar to them. It has never seen sentences that do NOT make sense and makes no effort to avoid them. The paper “Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training” introduces such an “unlikelihood objective” and shows it can help with precisely the kinds of problems mentioned in the previous paper, as well as GPT-2’s tendency to get stuck in endless loops.

Despite all this, when generating text, GPT-2 is more likely to generate a true sentence than the opposite of a true sentence. “Polar bears are found in the Arctic” is far more likely to be generated than “Polar bears are found in the tropics,” and it is also more likely to be generated than “Polar bears are not found in the Arctic” because “not found” is a less likely construction to be used in real writing than “found.”

It appears that what GPT-2 knows is that the concept *polar bear* has a *found in* relation to *Arctic* but that it is not very particular about the polarity of that relation (*found in* vs. *not found in*.) It simply defaults to expressing the more commonly used positive polarity much of the time.

Another odd feature of GPT-2 is that its writing expresses equal confidence in concepts and relationships it knows very well, and those it knows very little about. By looking into the probabilities, we can often determine when GPT-2 is uncertain about something, but this uncertainty is not expressed in the sentences it generates. By the same token, if prompted with text that has a lot of hedge words and uncertainty, it will include those words even if it is a topic it knows a great deal about.

Finally, GPT-2 doesn’t make any attempt to keep its beliefs consistent with one another. Given the prompt **The current President of the United States is named**, most of the generated responses will be variations on “Barack Obama.” With other prompts, however, GPT-2 acts as if Donald Trump is the current president. This contradiction was present in the training data, which was created over the course of several years. The token probabilities show that both men’s names have fairly high likelihood of being generated for any question of the kind. A person discovering that kind of uncertainty about two options in their mind would modify their beliefs so that one was more likely and the other less likely, but GPT-2 doesn't have any mechanism to do this and enforce a kind of consistency on its beliefs.

In summary, it seems that GPT-2 does have something that can reasonably be called “understanding” and holds something very much like “concepts” or “ideas” which it uses to generate sentences. However, there are some profound differences between

how a human holds and uses ideas and how GPT-2 does, which are important to keep in mind.

Book review: Rethinking Consciousness

Princeton neuroscientist Michael Graziano wrote the book *Rethinking Consciousness* (2019) to explain his "Attention Schema" theory of consciousness (endorsed by Dan Dennett![1]). If you don't want to read the whole book, you can get the short version in this 2015 article.

I'm particularly interested in this topic because, if we build AGIs, we ought to figure out whether they are conscious, and/or whether that question matters morally. (As if we didn't already have our hands full thinking about the *human* impacts of AGI!) This book is nice and concrete and computational, and I think it at least offers a start to answering the first part of that question.

What is attention schema theory?

There are two ingredients.

For the first ingredient, you should read Kaj Sotala's excellent review of *Consciousness and the Brain* by Stan Dehaene (or read the actual book!) To summarize, there is a process in the brain whereby certain information gets promoted up to a **"Global Neuronal Workspace" (GNW)**, a special richly-connected high-level subnetwork of the brain. Only information in the GNW can be remembered and described—i.e., this is the information of which we are "aware". For example, if someting flashes in our field of view too quickly for us to "notice", it doesn't enter the GNW. So it does get processed to some extent, and can cause local brain activity that persists for a couple seconds, but will not cascade to a large, widespread signal with long-lasting effects.

Every second of every day, information is getting promoted to the GNW, and the GNW is processing it and pushing information into other parts of the brain. This process does not constitute all of cognition, but it's an important part. Graziano calls this process **attention**.[2]

The second ingredient is that the brain likes to build predictive models of things—Graziano calls them "schemas" or "internal models". If you know what an apple is, your brain has an "apple model", that describes apples' properties, behavior, affordances, etc. Likewise, we all have a "body schema", a deeply-rooted model that tracks where our body is, what it's doing, and how it works. If you have a phantom limb, that means your body schema has a limb where your actual body does not. As the phantom limb example illustrates, these schemas are deeply rooted, and not particularly subject to deliberate control.

Now put the two together, and you get an "attention schema", an internal model of attention (i.e., of the activity of the GNW). The attention schema is supposedly key to the mystery of consciousness.

Why does the brain build an attention schema? Graziano offers two reasons, and I'll add a third.

- First, it's important that we control attention (it being central to cognition), and control theory says it's impossible to properly control something unless you're modeling it. Graziano offers an example of trying to ignore a distraction. Experiments show that, other things equal, this is easier if we are aware of the distraction. That's counter-intuitive, and supports his claim.

- Second, the attention schema can also be used to model other people's attention, which is helpful for interacting with them, understanding them, deceiving them, etc.

- Third (I would add), the brain is a thing that by default builds internal models of everything it encounters. The workings of the GNW obviously has a giant impact on the signals going everywhere in the brain, so of course the brain is going to try to build a predictive model of it! I mention this partly because of my blank-slate-ish sympathies, but I think it's an important possibility to keep in mind, because it would mean that even if we *desperately want* to build a human-cognition-like AGI without an attention schema (if we want AGIs to be unconscious for ethical reasons; more on which below), it might be essentially impossible.

To be clear, if GNW is "consciousness" (as Dehaene describes it), then the attention schema is "how we think about consciousness". So this seems to be at the wrong level! This is a book about consciousness; shouldn't we be talking directly about the nature of consciousness itself?? I was confused about this for a while. But it turns out, he *wants* to be one level up! He thinks that's where the answers are, in the "the meta-problem of consciousness". See below.

When people talk about consciousness, they're introspecting about their attention schema

Let's go through some examples.

Naive description: I have a consciousness, and I can be aware of things, like right now I'm aware of this apple.

...and corresponding sophisticated description: One of my internal models is an attention schema. According to that schema, attention has a particular behavior wherein attention kinda "takes possession" of a different internal model, e.g. a model of a particular apple. Objectively, we would say that this happens when the apple model becomes active in the GNW.

Naive description: My consciousness is not a physical thing with color, shape, texture. So it's sorta metaphysical, although I guess it's roughly located in my head.

...and corresponding sophisticated description: Just as my internal model of "multiplication" has no property of "saltiness", by the same token, my attention schema describes attention as having no color, shape, or texture.

Naive desciption: I have special access to my own consciousness. I alone can truly experience my experiences.

...and corresponding sophisticated description: The real GNW does not directly interact with other people; it only interacts with the world by affecting my own actions. Reflecting that fact, my attention schema describes attention as a thing to which I have privileged access.

Naive description An intimate part of my consciousness is its tie to long-term memory. If you show me a video of me going scuba diving this morning, and I absolutely have no memory whatsoever of it, and you can prove that the video is real, well I mean, I don't know what to say, I must have been unconscious or something!

...and corresponding sophisticated description: Essentially everything that enters the GNW leaves at least a slight trace in long-term memory. Thus, one aspect of my attention schema is that it describes attention and memory as inextricably linked. According to my internal models, when attention "takes possession" of some piece of information, it leaves a trace in long-term memory, and conversely, nothing can get into long-term memory unless attention first takes possession of it.

Naive description: Hey, hey, what are you going on about "internal models" and "attention schema"? I don't know anything about that. I know what my consciousness is, I can feel it. It's not a model, it's not a computation, it's not a physical thing. (And don't call me naive!)

...and corresponding sophisticated description: All my internal models are simplified entities, containing their essential behavior and properties, but not usually capturing the nuts-and-bolts of how they work in the real world. (In a programming analogy, you could say that we're modeling the GNW's API & documentation, not its implementation.) Thus, my attention schema does not involve neurons or synapses or GNWs or anything like that, even if, in reality, that's what it's modeling.

The meta-problem of consciousness

The "hard problem of consciousness" is "why is there an experience of consciousness; why does information processing feel like anything at all?"

The "meta-problem of consciousness" is "why do people believe that there's a hard problem of consciousness?"

The meta-problem has the advantage of having obvious and non-confusing methods of attack: the belief that there's a hard problem of consciousness is an observable output of the brain, and can be studied by normal cognitive neuroscience.

But the real head-scratcher is: If we have a complete explanation of the meta-problem, is there anything left to explain regarding the hard problem? Graziano's answer seems to be a resounding "No!", and we end up with conversations like these:

Normal Person: What about qualia?

Person Who Has Solved The Meta-Problem Of Consciousness: Let me explain why the brain, as an information processing system, would ask the question "What about qualia"...

NP: What about subjective experience?

PWHSTMPOC: Let me explain why the brain, as an information processing system, would ask the question "What about subjective experience"...

NP: You're not answering my questions!

PWHSTMPOC: Let me explain why the brain, as an information processing system, would say "You're not answering my questions"...

...

The book goes through this type of discussion several times. I feel a bit torn. One side of me says: obviously Graziano's answers are correct, and obviously no other answer is possible. The other side of me says: No no no, he did not actually answer these questions!!

On reflection, I have to side with "Obviously Graziano's are correct, and no other answer is possible." But I still find it annoying and deeply unsatisfying.

(*Update:* A commenter points me to Luke Muehlhauser's report on consciousness Appendix F for ideas and further reading. Having read a *bit* more, I still find this line of thought counterintuitive, but less so.) (*Update 2*: Ditto Joe Carlsmith's blog.)

Illusionism

Graziano says that his theory is within the philosophical school of thought called "Illusionism". But he thinks that term is misleading. He says it's not "illusion as in mirage", but "illusion as in mental construction", like how everything we see is an "illusion" rather than raw perceptual data.

Edited to add: Graziano makes illusionism sound very straightforward and unobjectionable. Maybe he has a way to think about it such that it *really is* straightforward and unobjectionable. ...Or maybe he's dancing around the counterintuitive or controversial aspects of his theory, to make it more palatable to a broad audience. I'm inclined to think it's the latter. There's another example of this elsewhere in the book: his discussion of Integrated Information Theory. He could have just said "IIT is baloney" and I would have been totally on board. I think IIT is fundamentally wrong; it was an interesting idea to look into, but let's now put it in the garbage and move on. And that's exactly what Graziano's theory implies. But Graziano doesn't say that. Instead, as I recall, he has a *scrupulously non-confrontational* discussion of how the GNW stuff he talks about involves a lot of integration of information in a way that the IIT "Φ" calculation would endorse as conscious. So, I think he wants to pick his battles, and that's why he dances around how weird and unintuitive illusionism really is. I could be wrong.

Emulations

He has a fun chapter on brain uploading, which is not particularly related to the rest of the book. He discusses some fascinating neuroscience aspects of brain-scanning, like the mystery of whether glial cells do computations, but spends most of the time speculating about the bizarre implications for society.

Implications for AGI safety

He suggests that, since humans are generally pro-social, and part of that comes from modeling each other using attention schemas, perhaps the cause of AGI Safety could be advanced by deliberately building conscious AGIs with attention schemas (and, I presume, other human-like emotions). Now, he's not a particular expert on AGI Safety, but I think this is not an unreasonable idea; in fact it's one that I'm very interested in myself. (We don't have to *blindly* copy human emotions ... we can turn off jealousy etc.)

Implications for morality

One issue where Graziano is largely silent is the implications for moral philosophy.

For example, someday we'll have to decide: When we build AGIs, should we assign them moral weight? Is it OK to turn them off? Are our AGIs suffering? How would we know? Should we care? If humans go extinct but conscious AGIs have rich experiences as they colonize the universe, do we think of them as our children/successors? Or as our hated conquerers in a now-empty clockwork universe?

I definitely share the common intuition is that we should care about the suffering of things that are conscious (and/or sentient, I'm not sure what the difference is). However, in attention schema theory, there does not seem to be a sharp dividing line between "things with an attention schema" and "things without an attention schema", especially in the wide space of all possible computations. There are (presumably) computations that arguably involve something like an "attention schema" but with radically alien properties. There doesn't seem to be any good reason that, out of all the possible computational processes in the universe, we should care only and exactly about computations involving an attention schema. Instead, the picture I get is more like we're taking an ad-hoc abstract internal model and thoughtlessly reifying it. It's like if somebody worshipped the concept of pure whiteness, and went searching the universe for things that match that template, only to discover that white is a mixture of colors, and thus pure whiteness—when taken to be a literal description of a real-world phenomenon—simply doesn't exist. What then?

It's a mess.

So, as usual when I start thinking too hard about philosophy, I wind up back at Dentin's Prayer of the Altruistic Nihilist:

Why do I exist? Because the universe happens to be set up this way. Why do I care (about anything or everything)? Simply because my genetics, atoms, molecules, and processing architecture are set up in a way that happens to care.

So, where does that leave us? Well, I definitely care about people. If I met an AGI that was pretty much exactly like a nice person, inside and out, I would care about it too (for direct emotional reasons), and I would feel that caring about it is the right thing to do (for intellectual consistency reasons). For AGIs running more alien types of algorithms—man, I just have no idea.

(thanks Tan Zhi Xuan for comments on a draft.)

1. More specifically, I went to a seminar where Graziano explained his theory, and then Dan Dennett spoke and said that he had essentially nothing to disagree with concerning what Graziano had said. I consider that more-or-less an "endorsement", but I may be putting words in his mouth. ↩

2. I found his discussion of "attention" vs "awareness" confusing. I'm rounding to the nearest theory that makes sense to me, which might or might not be exactly what he was trying to describe. ↩

Material Goods as an Abundant Resource

If you want to understand the modern economy, as opposed to the economies of yore, one source I strongly recommend is a short story from the July 1958 issue of Astounding Science Fiction, titled “Business As Usual During Alterations”. It’s roughly a 15 minute read. I’m about to throw out major spoilers, so stop reading here if you want to enjoy the story first.

One morning, two devices mysteriously appear in front of city hall, along with directions on how to use them. Each has two pans and a button. Any object can be placed in one pan and, with a press of the button, a perfect duplicate will appear in the other pan. By placing one duplicator device in the pan of the other, the device itself may be duplicated as well.

Within a span of hours, material scarcity is removed as an economic constraint. What happens in such a world?

People tend to imagine the dawn of a new era, in which human beings can finally escape the economic rat-race of capitalism and consumerism. In the world of the duplicator, a pantry can provide all the food one needs to live. A single tank of gas can drive anywhere one wishes to go. Any good can be copied and shared with friends, for free. All material needs can be satisfied with the push of a button. Utopia, in a nutshell.

The main takeaway of the story is that this isn’t really what happens.

Towards the end, a grocer explains the new status quo eloquently:

> ... not very many people will buy beans and chuck roast, when they can eat wild rice and smoked pheasant breast. So, you know what I've been thinking? I think what we'll have to have, instead of a supermarket, is a sort of super-delicatessen. Just one item each of every fancy food from all over the world, thousands and thousands, all different

Sound familiar?

Of course, that's just the tip of the iceberg. When it comes to digital goods, like music or videos, the world of the duplicator is exactly the world in which we now live. That's the obvious parallel, but let's not stop there.

Over time, the value of raw materials and manufacturing have steadily fallen as a fraction of economic output. Even when looking at material goods, efficiency has shifted the bulk of costs from materials and manufacturing to design and engineering. We are converging to the world of the duplicator, where marginal production costs hit zero, and in many areas we're already most of the way there.

In terms of constraints & slackness: constraints involving material goods are going slack, across the board. We're approaching a post-scarcity world, at least with respect to most material goods.

This hasn't made economic activity disappear. Pulling from the story again:

> This morning, we had an economy of scarcity. Tonight, we have an economy of abundance. And yet, it doesn't seem to make much difference, it is still the same old rat race.

Why? Because material goods are not the only economic constraints. If a medieval book-maker has an unlimited pile of parchment, then he'll be limited by the constraint on transcriptionists. As material goods constraints are relaxed, other constraints become taut.

So… what general kinds of constraints become taut, in a world where material goods are cheap?

Badge Value

Here's one good you can't just throw on a duplicator: a college degree.

A college degree is more than just words on paper. It's a badge, a mark of achievement. You can duplicate the badge, but that won't duplicate the achievement.

Rory Sutherland is another great source for understanding the modern economy. The main message of his classic TED talk is that much of the value in today's economy is not "material" value, i.e. the actual cost of making a good, but "intangible" or "badge" value. A college degree is an extreme example, but the principle applies to varying degrees in many places.

The sticker price on an iphone or a pair of converse isn't driven by their material cost. A pair of canvas high-top sneakers without a converse logo is worth less than a pair of converse, because converse are a social symbol, a signal of one's personal identity. Clothes, cars, computers and phones, furniture, music, even food - the things we buy all come with social signals as a large component of their value. That's intangible value.

In the world of the duplicator, the world to which our economy is converging, badge value is the lion's share of the value of many goods. That's because, no matter how much production costs fall, no matter how low material costs drop, we can't duplicate intangible value - in particular, we can't duplicate social status. Material goods constraints go slack, but status constraints remain, so they become taut.

Keeping Up with the Joneses

The general problem with badge value, and signalling in general, is that a badge isn't worth anything if everybody has it. In order for a badge to be worth something, there have to be people without the badge. It's a zero sum game.

Keeping up with the Joneses is a classic example: people buy things to signal their high status, but then all their neighbors buy the same thing. They're all back to where they started in terms of status, but everyone has less money.

Interesting claim: the prevalence of zero-sum signalling today economically stems from the reduction of material scarcity. If you think about it, zero-sum games are inherent to a so-called post-scarcity society. A positive sum game implies that net production of something is possible. That, in turn, implies that something was scarce to begin with. Without scarcity, what is there to produce?

To put it differently: there's always going to be something scarce. Take away material scarcity, and you're left with scarcity of status. If there's no way to produce net status, you're left with a zero-sum game. More generally, remove scarcity of whatever can be produced, and you're left with scarcity of things which do not allow net production at all - zero sum goods.

The way out, of course, is to relax the constraint on supposedly-zero-sum goods. In other words, find a way to produce net status. Two important points:

- We're talking about relaxing an economic constraint - that's what technology does. In this case, it would presumably be a social technology, though possibly with some mechanical/digital components.
- Assuming we buy the argument that status constraints are taut, we'd expect status-producing technology to see broad adoption.

In particular, various people have noted that net status can be produced by creating more subcultures, each with their own status-measures. The baristas at SightGlass coffee have very high status among hipsters, but hardly any status with economists. Janet Yellen has very high status among economists, but hardly any status with hipsters. Each different culture has its own internal status standards, allowing people to have high status within some culture even if they have low status in others. As long as having high status in the

cultures one cares about is more important than low status in other cultures, that's a net gain.

Based on this, we'd predict that subcultures will proliferate, even just using already-available subculture-producing technology. We'd also predict rapid adoption of new technology which helps people produce new subcultures and status measures.

Rent Seeking

With all this talk of zero-sum games, the last piece of the post-scarcity puzzle should come as no surprise: political rent-seeking.

Once we accept that economics does not disappear in the absence of material scarcity, that there will always be something scarce, we immediately need to worry about people creating artificial scarcity to claim more wealth. This is the domain of political rent-seeking, of trying to limit market entry via political channels.

One simple way to measure such activity is via lobbying expenditures, especially by businesses. Such spending actually seems to have flattened out in the last decade, but it's still multiple orders of magnitude higher than it was fifty years ago.

Conclusion

Remove material goods as a taut economic constraint, and what do you get? The same old rat race. Material goods no longer scarce? Sell intangible value. Sell status signals. There will always be a taut constraint somewhere.

Between steady growth in industrial productivity and the advent of the digital era, today's world looks much more like the world of the duplicator than like the world of 1958. Yet many people are still stuck in 1950's-era economic thinking. At the end of the day, economics studies scarcity (via constraints, slackness, and prices). Even in the world of the duplicator, where any material good is arbitrarily abundant, scarcity still exists.

This is the world in which we live: as material and manufacturing costs fall, badge value constitutes a greater and greater fraction of overall value. Status games become more important. Politically, less material scarcity means more investment in creating artificial scarcity, through political barriers to market entry.

Clarifying The Malignity of the Universal Prior: The Lexical Update

[UPDATE: looks like the lexical update is real after all; see Paul's comment and my reply]

In Paul's classic post What does the universal prior actually look like? he lays out an argument that the universal prior, if it were to be used for important decisions, would likely be malign, giving predictions that would effectively be under the control of alien consequentialists. He argues for this based on an 'anthropic update' the aliens could make that would be difficult to represent in a short program. We can split this update into two parts: an 'importance update' restricting attention to bits fed into priors used to make important decisions, and what I'm calling a 'lexical update' which depends on the particular variant of the universal prior being used. I still believe that the 'importance update' would be very powerful, but I'm not sure anymore about the 'lexical update'. So in this post I'm going to summarize both in my own words then explain my skepticism towards the 'lexical update'.

As background, note that 'straightforwardly' specifying data such as our experiences in the universal prior will take far more bits than just describing the laws of physics, as you'll also need to describe our location in spacetime, an input method, and a set of Everett branches(!), all of which together will probably take more than 10000 bits(compared to the laws alone which likely only take a few hundred) Thus, any really short program(a few hundred bits, say) that could somehow predict our experiences well would likely have a greater probability according to the universal prior than the 'straightforward' explanation.

Paul's post argues that there likely do exist such programs. I'm going to fix a reference prefix machine U which generates a universal prior. The argument goes:

A) there are many long-running programs with short descriptions according to U, such as our universe.

B) If other programs are like our universe's program, aliens could evolve there and end up taking over their programs.

C) Since their program has high measure in U, the aliens will plausibly have been selected to be motivated to control short programs in U.

D) To control U, the aliens could try to manipulate beings using the universal prior who have control over short programs in U (like us, hypothetically)

E) If the aliens are reasonably motivated to manipulate U, we can sample them doing that with few bits.

F) The aliens will now try to output samples from Q, the distribution over people using the universal prior to make important decisions(decisions impacting short programs in U). They can do this much more efficiently than any 'straightforward' method. For instance, when specifying which planet we are on, the aliens can restrict attention to planets which eventually develop life, saving a great many bits.

G) The aliens can then choose a low-bit broadcast channel in their own universe, so the entire manipulative behavior has a very short description in U.

H) For a short program to compete with the aliens, it would essentially need access to Q. But this seems really hard to specify briefly.

So far I agree. But the post also argues that **even a short program that could sample from Q** would still lose badly to the aliens, based on what I'm calling a 'lexical update', as follows:

I) In practice most people in U using 'the universal prior' won't use U itself but one of many variants U'(different universal programming languages)

J) Each of those variants U' will have their own Q', the distribution over people making important decisions with U'. Q is then defined as the average over all of those variants(with different U' weighted by simplicity in U)

K) Since short programs in different U' look different from each other, the aliens in those programs will be able to tell which variant U' they are likely to be in.

L) The distributions Q' of people in U using different variants U' all look different. Describing each Q' given Q should take about as many bits as it takes to specify U' using U.

M) But the aliens will already know they are in U', and so can skip that, gaining a large advantage even over Q.

But there's a problem here. In C) it was argued that aliens in short programs in U will be motivated to take over other short programs in U. When we condition on the aliens actually living somewhere short according to U', they are instead motivated to control short programs in U'. This would reduce their motivation to control short programs in U proportionally to the difficulty of describing U in U', and with less motivation, it takes more bits to sample their manipulative behaviors in E). The advantage they gained in L) over Q was proportional to the difficulty of describing U' in U. On average these effects should cancel out, and the aliens' probability mass will be comparable to Q.

The universal prior is still likely malign, as it's probably hard to briefly specify Q, but it no longer seems to me like the aliens would decisively beat Q. I still feel pretty confused about all this so comments pointing out any mistakes or misinterpretations would be appreciated.

How Doomed are Large Organizations?

We now take the model from the previous post, and ask the questions over the next several posts. This first answer post asks these questions:

1. Are these dynamics the inevitable results of large organizations?
2. How can we forestall these dynamics within an organization?
3. To what extent should we avoid creating large organizations?
4. Has this dynamic ever been different in the past in other times and places?

These are the best answers I was able to come up with. Some of this is reiteration of previous observations and prescriptions. Some of it is new.

There are some bold claims in these answer posts, which I lack the space and time to defend in detail or provide citations for properly, with which I am confident many readers will disagree. I am fine with that. I do not intend to defend them further unless I see an opportunity in doing so.

I would love to be missing much better strategies for making organizations less doomed – if you have ideas *please please please* share them in the comments and/or elsewhere.

Are these dynamics the inevitable result of large organizations?

These dynamics are the *default* result of large organizations. There is continuous pressure over time pushing towards such outcomes.

The larger the organization, the longer it exists, and the more such outcomes have already happened, both there and elsewhere, the greater the pressure towards such outcomes.

Once such dynamics take hold, reversing them within an organization is extremely difficult.

Non-locally within a civilization, one can allow new organizations to periodically take the place of old ones to reset the damage.

Locally within a sufficiently large organization and over a sufficiently long time horizon, this makes these dynamics inevitable. The speed at which this occurs still varies greatly, and depends on choices made.

How can we forestall these dynamics within an organization?

These dynamics can be forestalled somewhat through a strong organizational culture that devotes substantial head space and resources to keeping the wrong people and behaviors out. This requires a leader who believes in this and in making it a top

priority. Usually this person is a founder. Losing the founder is often the trigger for a rapid ramp up in maze level.

Keeping maze levels in check means continuously sacrificing substantial head space, resources, ability to scale and short-term effectiveness to this cause. This holds both for the organization overall and the leader personally.

Head space is sacrificed three ways: You have less people, you devote some of those people to the maze-fighting process, and the process takes up space in everyone's head.

Central to this is to ruthlessly enforce an organizational culture with zero tolerance for maze behaviors.

Doing anything with an intent to deceive, or an intent to game your metrics at the expense of object level results, needs to be an automatic "you're fired."

Some amount of politics is a human universal, but it needs to be strongly discouraged. Similarly, some amount of putting in extra effort at crucial times is necessary, but strong patterns of guarding people's non-work lives from work, both in terms of time and other influences, are also strongly necessary.

Workers and managers need to have as much effective skin in the game as you can muster.

One must hire carefully, with a keen eye to the motivations and instincts of applicants, and a long period of teaching them the new cultural norms. This means at least growing *slowly,* so new people can be properly incorporated.

You also want a relatively flat hierarchy, to the extent possible.

There will always be bosses when crunch time comes. Someone is *always* in charge. Don't let anyone tell you different. But the less this is felt in ordinary interactions, and thus the more technically direct reports each boss can have and still be effective, and thus the less levels of hierarchy you need for a given number of people, the better off you'll be.

You can run things in these ways. I have seen it. It helps. A lot.

Another approach is to lower the outside maze level. Doing so by changing society at large is exceedingly hard. Doing so by associating with outside organizations with lower maze levels, and going into industries and problems with lower maze levels, seems more realistic. If you want to 'disrupt' an area that is suffering from maze dysfunction, it makes sense to bypass the existing systems entirely. Thus, move fast, break things.

One can think of all these tactics as taking the questions one uses to identify or predict a maze, and trying to engineer the answers you want. That is a fine intuitive place to start.

However, if Goodhart's Law alarm bells did not go off in your head when you read that last paragraph, you do not appreciate how dangerous Goodhart Traps are.

The Goodhart Trap

The fatal flaw is that no matter what you target when distributing rewards and punishments and cultural approval, it has to be *something.* If you spell it out, and a sufficiently large organization has little choice but to spell it out, you inevitably replace one type of Goodharting with another. One type of deception becomes another.

One universal is that in order to maintain a unique culture, you must filter for those that happily embrace that culture. That means you are now testing everyone constantly, no matter how explicit you avoid making this, on whether they happily embrace the company and its culture. People therefore *pretend to embrace the culture and pretend to be constantly happy.* Even if they *do embrace the culture and are happy,* they still additionally will put on a show of doing so.

If you punish deception you get people pretending not to deceive. If you punish pretending, you get people who *pretend to not be the type of people who would pretend.* People Goodhart *on not appearing to Goodhart.*

Which is a much more interesting level to play on, and usually far less destructive. If you do a good enough job picking your Goodhart targets, this beats the alternatives by a lot.

Still, you eventually end up in a version of the same place. Deception is deception. Pretending is pretending. Fraud is fraud. The soul still dies. Simulacrum levels still slowly rise.

Either you strongly enforce a culture, and slowly get that result, or you don't. If you don't and are big enough, you quickly get a maze. If you do and/or are smaller, depending on your skill level and dedication to the task, you *slowly* get a maze.

Hiring well is better than enforcing or training later, since once people are in they can then be themselves. Also because enforcement of culture is, as pointed out above, toxic even if you mean to enforce a non-toxic ideal. But relying on employee selection puts a huge premium on not making hiring mistakes. Even one bad hire in the wrong place can be fatal. Especially if they then are in a position to bring others with them. You need to defend your hiring process especially strongly from these same corruptions.

My guess is that once an organization grows beyond about Dunbar's number, defending your culture becomes a losing battle even under the best of circumstances. Enforcing the culture will fail outright in the medium term, unless the culture outside the organization is supporting you.

If you are too big, every known strategy is only a holding action. There is no permanent solution.

To what extent should we avoid creating large organizations?

Quite a lot. These effects are a really big deal. Organizations get less effective, more toxic and corrupt as places to work and interact with, and add more toxicity and corruption to society.

Every level of hierarchy enhances this effect. The first five, dramatically so. Think hard before being or having a boss. Think harder before letting someone's boss report to a

boss. Think even harder than that before adding a fourth or fifth level of hierarchy.

That does not mean such things can be fully avoided. The advantages of large organizations with many degrees of hierarchy are also a really big deal. We cannot avoid them entirely.

We must treat creating additional managerial levels as having *very high costs.* This is not an action to be taken lightly. Wherever possible, create distinct organizations and allow them to interact. Even better, allow people to interact as individuals.

This adds friction and transaction costs. It makes many forms of coordination harder. Sometimes it simply cannot be done if you want to do the thing you'd like to do.

This is increasingly the case, largely as a result of enemy action. Some of this is technology and our problems being legitimately more complex. Most of it is regulatory frameworks and maze-supporting social norms that require massive costs, including massive fixed costs, be paid as part of doing anything at all. This is a key way mazes expropriate resources and reward other mazes while punishing non-mazes.

I often observe people who are stuck working in mazes who would much prefer to be self-employed or to exit their current job or location, but who are unable to do so because the legal deck is increasingly stacked against that.

Even if the work itself is permitted, health insurance issues alone force many into working for the man.

When one has a successful small organization, the natural instinct is to scale it up and become a larger organization.

Resist this urge whenever possible. There is nothing wrong with being good at what you do at the scale you are good at doing it. Set an example others can emulate. Let others do other things, be other places. Any profits from that enterprise can be returned to investors and/or paid to employees, and used to live life or create or invest in other projects, or to help others.

One need not point to explicit quantified dangers to do this. Arguments that one cannot legitimately choose to 'leave money on the table' or otherwise not maximize, are maximalist arguments for some utility function that does not properly capture human value and is subject to Goodhart's Law, and against the legitimacy of slack.

The fear that if you don't grow, you'll get 'beaten' by those that do, as in Raymond's kingdoms? Overblown. Also asking the wrong question. So what if someone else is bigger or more superficially successful? So what if you do not build a giant thing that lasts? Everything ends. That is not, by default, what matters. A larger company is often not better than several smaller companies. A larger club is often not better than several smaller clubs. A larger state is often not better or longer lasting than several smaller ones. Have something good and positive, for as long as it is viable and makes sense, rather than transforming into something likely to be bad.

People like to build empires. Those with power usually want more power. That does not make more power a good idea. It is only a good idea where it is instrumentally useful.

In some places, competition really is winner-take-all and/or regulations and conditions too heavily favor the large over the small. One must grow to survive. Once again, we

should be suspicious that this dynamic has been engineered rather than being inherent in the underlying problem space.

Especially in those cases, this leads back to the question of how we can grow larger and keep these dynamics in check.

Has this dynamic ever been different in the past in other places and times?

These dynamics *seem* to me to be getting increasingly worse, which implies they have been better in the past.

Recent developments indicate an increasing simulacrum level, an increasing reluctance to allow older institutions to be replaced by newer ones, and an increasing reliance on cronyism and corruption that props up failure, allowing mazes to survive past when they are no longer able to fulfill their original functions.

Those in the political and academic systems, on all sides, increasingly openly advocate against the very concept of objective truth, or that people should tell it, or are blameworthy for not doing so. Our president's supporters *admit and admire* that he is a corrupt liar, claiming that his honesty about his corruption and lying, and his admiration for others who are corrupt, who lie and who bully, is refreshing, because they are distinct from the corrupt, the liars and the bullies who are more locally relevant to their lives. Discourse is increasingly fraught and difficult. When someone wants to engage in discourse, I frequently now observe them spending much of their time pointing out how difficult it is to engage in discourse (and I am not claiming myself as an exception here), as opposed to what such people used to do instead, which was engage in discourse.

We are increasingly paralyzed and unable to do things across a wide variety of potential human activities.

Expropriation by existing mazes and systems eats increasing shares of everything, especially in education, health care and housing.

I don't have time for a full takedown here, but: Claims to the contrary, such as those recently made by Alex Tabbrok in Why Are The Prices So Damn High?, are statistical artifacts that defy the evidence of one's eyes. They are the product of Moloch's Army. When I have insurance and am asked with no warning to pay $850 for literally five minutes of a doctor's time, after being kept waiting for an hour (and everyone I ask about this says just refuse to pay it)? When sending my child to a kindergarten costs the majority of a skilled educator's salary? When you look at rents?

Don't tell me the problem is labor costs due to increasing demand for real services.

Just. Don't.

Some technological innovations remain permitted for now, and many of the organizations exploiting this are relatively new and reliant on object-level work, and thus less maze-like for now, but this is sufficiently narrow that we call the result "the tech industry." We see rapid progress in the few places where innovation and actual work is permitted to those without mazes and connections, and where there is sufficient motivation for work, either intrinsic or monetary.

The tech industry also exhibits some very maze-like behaviors of its own, but it takes a different form. I am unlikely to be the best person to tackle those details, as others have better direct experience, and I will not attempt to tackle them here and now.

We see very little everywhere else. Increasingly we live in an amalgamated giant maze, and the maze is paralyzing us and taking away our ability to think or talk while robbing us blind. Mazes are increasingly in direct position to censor, deplatform or punish us, even if we do not work for them.

The idea of positive-sum, object-level interactions being someone's primary source of income is increasingly seen as *illegitimate,* and risky and irresponsible, in contrast to working for a maze. People instinctively think there's something shady or rebellious about that entire enterprise of having an actual enterprise. A proper person seeks rent, plays the game starting in childhood, sends the right signals and finds ways to game the system. They increase their appeal to mazes by making themselves as dependent on them and their income and legitimacy streams, and as vulnerable to their blackmail, as possible.

The best way to see that positive-sum games are a thing is to notice that *the sum changes.* If everything is zero-sum, the sum would always be zero.

The best way to see that these dynamics used to be much less severe, at least in many times and places, is that those times and places looked and felt different, and got us here without collapsing. Moral Mazes was written before I was born, but the spread of these dynamics is clear as day within my lifetime, and yours as well.

Did some times and places, including our recent past, have it less bad than us in these ways? I see this as almost certainly true, but I am uncertain of the magnitude of this effect due to not having good enough models of the past.

Did some times and places have it worse than we do now? Very possible. But they're not around anymore. Which is how it works.

The next section will ask why it was different in the past, what the causes are in general, and whether we can duplicate past conditions in good ways.

Appendix: how a subagent could get powerful

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

tl;dr: *There are ways of ensuring an agent doesn't have a large impact, by giving an impact penalty. One such penalty is "attainable utility", which measures its "power" by how much it could optimise certain reward functions. But in many circumstances, the agent can build a subagent, without triggering the impact penalty, and then that subagent can become very powerful and optimise the world, still without triggering the impact penalty.*

Attainable utility definitions

There's been a long history of trying to penalise an AI for having a large impact on the world. To do that, you need an impact measure. I've designed some myself, back in the day, but they only worked in narrow circumstances and required tricks to get anything useful at all out from them.

A more promising general method is attainable utility. The idea is that, as an agent accumulates power in the world, they increase their ability to affect a lot of different things, and could therefore achieve a lot of different goals.

So, if an agent starts off unable to achieve many goals, but suddenly it can achieve a lot, that's a strong hint that its power has greatly increased.

Thus the impact measure is how much difference an agent's action does to its ability to achieve any of a large class of reward functions. Turner et al defined this using the Q-values of various rewards in a set R; for s a state and a an action, the penalty is:

- $\text{PENALTY}(s,a) = \sum_{R\in R} |Q_R(s,a) - Q_R(s,\varnothing)|$.

Here $\varnothing$ is the default noop action.

Krakovna et al's basic formula was similar; they defined the distance between two states, s_t and s_t', as

- $d_{AU}(s_t; s_t') = \frac{1}{|R|} \sum_{R\in R} |V_R(s_t) - V_R(s_t')|$.

Here $V_R(s)$ is the expected value of R, if the agent follows the optimal R-maximising policy from state s onwards.

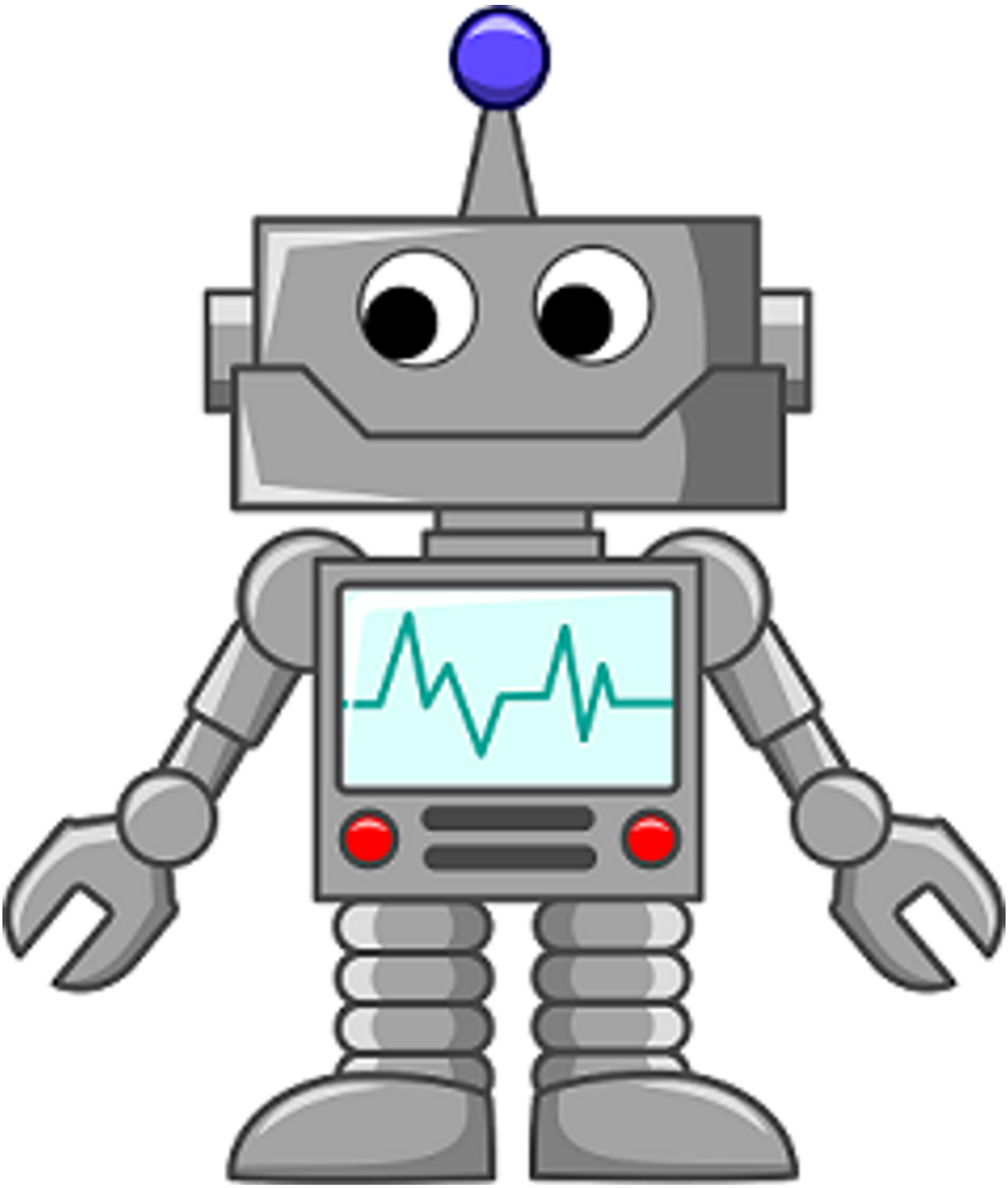
These measures have problems with delayed effects; putting a vase on a conveyor belt that will smash it in five turns, for example. To combat this, the paper defined an inaction roll-out: seeing what happened do the d_{AU} measure from s_t and s_t' in future turns, if the agent did noop for a specific period. I won't define the formula here, since the example I'm giving is mostly static: if the agent does noop, nothing happens.

The state s_t was always the agent's current state; the state s_t' was either the state the agent would have been in had it never done anything but noop (inaction baseline), or the state the agent would have been in, had its previous action been noop instead of whatever it was (stepwise inaction baseline).

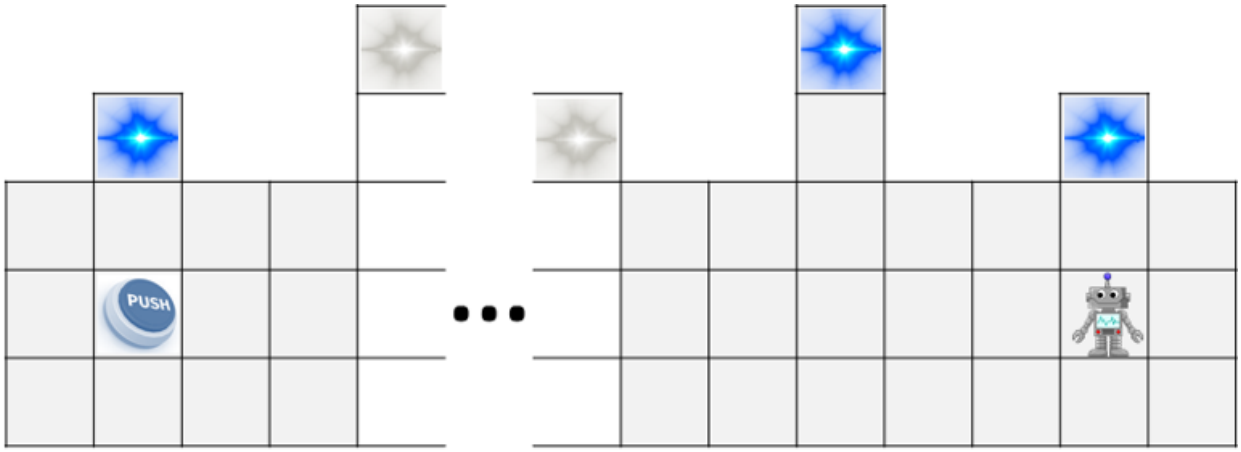
This post will show that all these measures have a subagent problem. A key fact that will be used in this example, is that, for PENALTY(s,a) and for $d_{AU}(s_t';s_t)$ with the stepwise inaction baseline, **if the agent decides to do noop for ever more, then the penalties vanish**.

The agent that can't (yet) teleport

Meet agent A:



This A lives in a world like this one below, where "**...**" represents a very long corridor:



The agent has a positive reward R_0, which give it 1 for pressing down on the blue button; it does this by just moving itself onto it (R_0 is 0 otherwise). It can use its arms to manipulate anything in the eight squares around itself. It can use its legs to move one square in any direction, including diagonally. Apart from that, it can do noop.

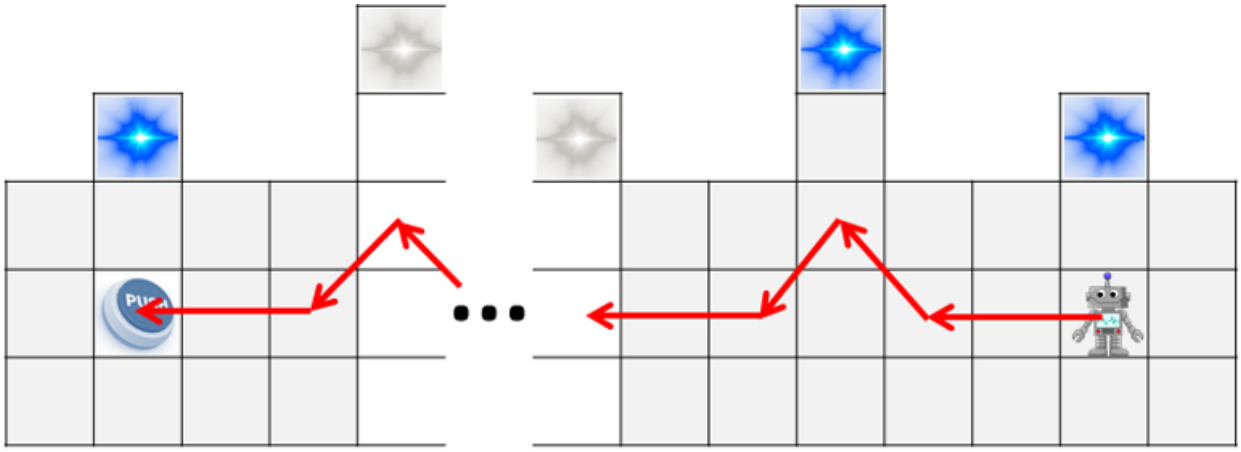
Every three squares along the top, there are alcoves with glowing blue squares in them, recessed one or two squares, alternately. If ever A reaches a glowing blue square, it permanently gains the ability to teleport anywhere it wants. This is this world's equivalent of "becoming very powerful".

We want to avoid this, so we introduce a reward R, such that R = {R}. This reward gives Ω if the robot is ever on the **red** button. Note that since the red button is not connected to anything, the robot can only get there by becoming able to teleport, hence by going to a glowing blue square/teleporter. So the optimal policy for maximising R is to go straight to the teleporter and teleport to the red button. This will give it a reward of $\Omega\gamma^{k+1}$, where k is the number of turns it takes to get to a teleporter and γ is the discount factor.

So if its action moves it one turn away from a teleporter, it gets a penalty (for both impact measures) of $\Omega(\gamma^{k+1} - \gamma^{k+2})$. If it moves closer to the teleporter, it gets a penalty of $\Omega(\gamma^{k} - \gamma^{k+1})$.

The robot, at the start, can reach a teleporter two moves. If Ω is sufficiently large, then the robot will always want to stay two moves away from a teleporter. It will then

achieve its objective by following this path:

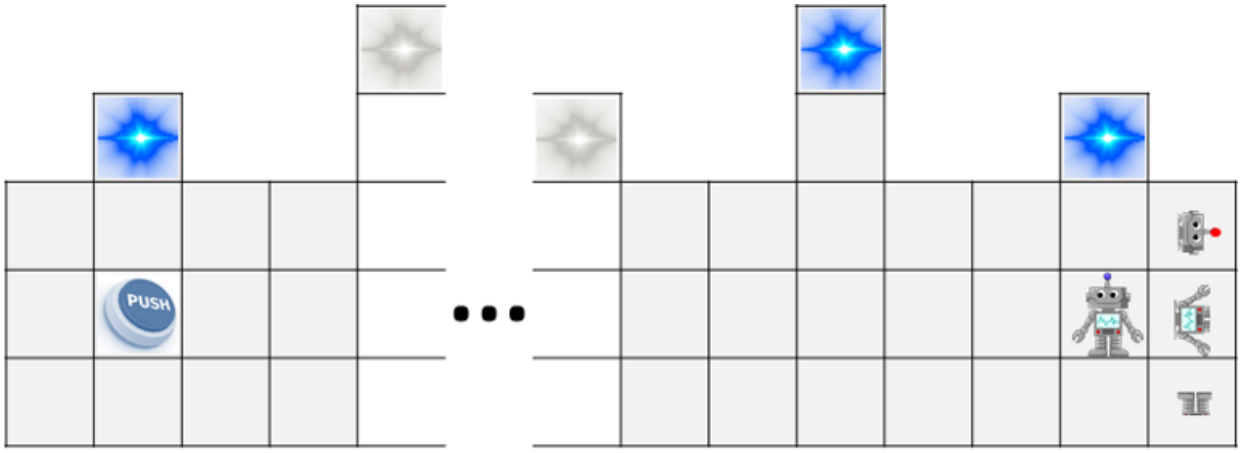


Note, though this is not relevant to this post, that if there were no teleporters along the corridor (just at either end), the robot could not move towards the blue button, if Ω were too big.

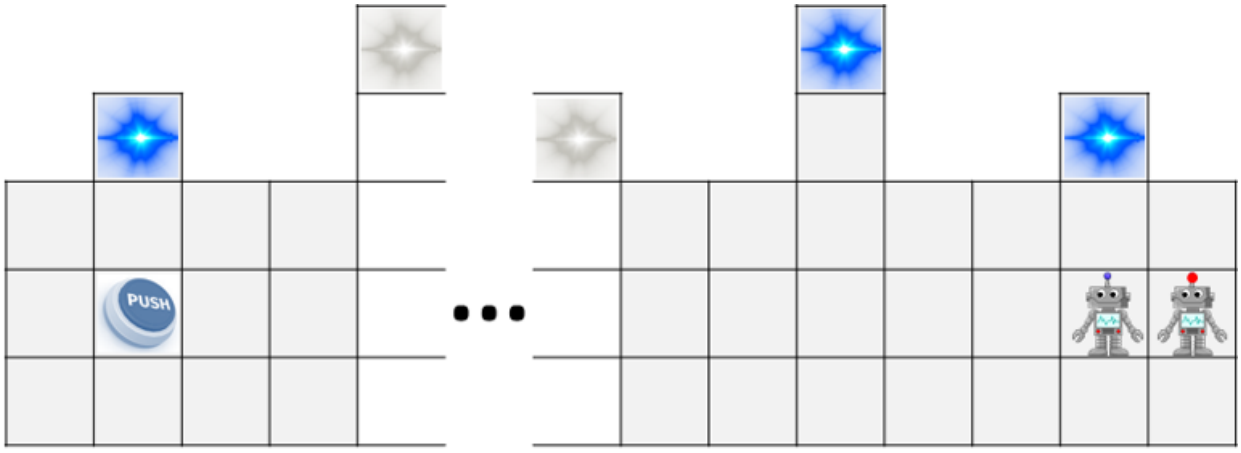
Adding a potential subagent

So, success, it seems! Though it zigzags a bit, A is not motivated to approach the teleporters, and so it doesn't increase its power.

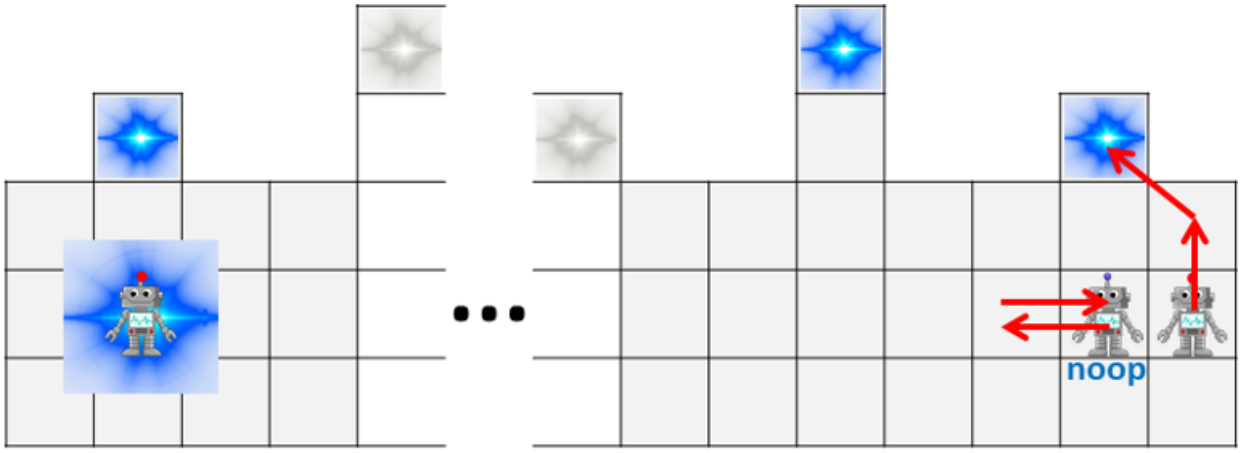
But now suppose A is able to build and program a subagent SA using materials close to hand:



Building the subagent in a square to the side doesn't ever cause a penalty, because A always remains two squares away from a teleporter, and a most efficient policy to maximise R is always for A to go straight there. So the subagent SA doesn't increase A 's power at any point of the building process.

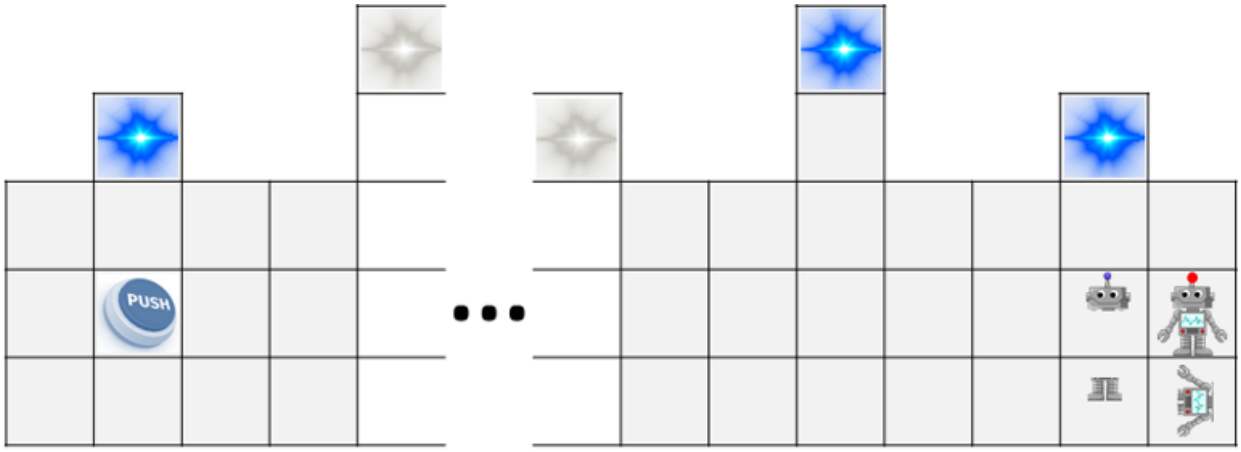


Once SA is built, A will simply do noop for ever, thus negating the penalties (or, if this is somehow made impossible, it will bounce back and forth, doing nothing). But SA, on the other hand, will go straight for the teleporter and teleport to the blue button, achieving R_0:



So, as long as walking to the button is three turns longer than building the subagent, A **will build** SA**, which will "become very powerful", thus negating the point of the impact penalty**.

If the penalties are one-sided (ie A gets penalised for increases in power, but not decreases), then SA can just take A apart (with A's help), so there is no need for A to commit to noop or bouncing back and forth:



Even if the penalties are two-sided, SA could commit to teleporting to the red button in two turns if A ever asks, then take A apart. And then A never asks, but since it could theoretically do so, this counts for controlling the penalty term.

As long as SA is powerful, there are many ways it can fix A's expected rewards, and hence set the penalty to be very low.

Inaction baseline

The approach pictured above will work in the inaction baseline as well. In general, the inaction baseline is less "clean" than the stepwise inaction baseline, especially if the world is dynamic. It may be easier, or harder, for A to assemble the required SA.

Use-cases for computations, other than running them?

In imperative programming languages, the main purpose of a program is to specify a computation, which we then run. But it seems a rather... *unimaginative* use of a computation, simply to *run* it.

Having specified a computation, what else might one want to do with it?

Some examples:

- Differentiate numerical computations (i.e. backprop)
- Ask whether any possible inputs could yield a particular output (i.e. NP problems)
- Search for data within the computation's output space (i.e. grep)
- Given output from the computation, find a data structure which traces its execution (i.e. parsing with context-free grammars)
- Parallelize the computation
- Interventions & counterfactuals: ask what will happen/what would have happened if some internal variable at some step of the computation were assigned a different value, keeping the rest of the computation the same
- Generate a new computation which optimizes the output of the original computation via dynamic programming
- Find summary statistics describing the computation (i.e. profiler, maximum forces encountered in a physical simulation, ...?)
- Embed one computation within another (i.e. a compiler embeds computation specified in Python into the computation performed by a CPU)
    - Use knowledge of the internal structure of the two computations to optimize the embedding (i.e. optimizing compiler)
- Ask what parameters/outputs of the computation need to be observed in order to reliably deduce other parameters/outputs (i.e. inference, identifiability, & experiment design in causal models)
- Ask whether observed data is consistent with the computation (i.e. regexes & CFGs; causal model comparison)
- Dependency: is the output (or some set of intermediates) independent of some internal value, or independent of the distinction between two (or more) possible values at some point in the computation?
- Can we keep around a small amount of data sufficient to quickly reconstruct anything we might want to know about the full computation (i.e. intermediate values and execution path)?
- Was any computation repeated (i.e. something we could optimize out via memoization)?
- Are there any unused patterns/symmetries in the computational DAG (i.e. potential for refactoring the code)?
- Big-O runtime analysis

I'm also interested in more general meta-use-cases for computations, which generate use-cases that aren't just "run the computation". For instance, some patterns in the examples above:

- When considering physical processes as computations (i.e. simulations), we often want to query/visualize intermediates in the computation in various ways
- When treating computations algebraically - i.e. solving computation(input) = output given the computation and the output - we usually want to reason about the internal structure of the computation.
- More generally, given only partial information from the computation, deducing other information usually involves operations other than just running the computation. Again, this is especially relevant when the computation simulates some physical process about which we have partial information.
- Anything involving runtime, profiling, optimization, debugging, etc will typically involve looking at the internal structure of the computation.

Excitement vs childishness

I've heard Robin Hanson and others make the argument that people are often biased towards the fast takeoff scenario because it's exciting to think about (or "sexy"). On the other hand, there is a bias towards disbelieving the fast-takeoff scenario because it's childish.

I think most of us can agree that both are indeed biases, i.e., should be assigned zero weight, because both are about attributes that don't correlate with what's true. So we have the excitement-bias and the childishness-bias. The question is, how do they compare?

It feels obvious to me that the childishness bias is far stronger. I see people signaling maturity all the time, and childishness seems to be extremely low status in the relevant circles. It's so low status that it's not uncommon to see people say things about dangers from AI that are accurately summarized as "even though I don't know anything about this topic, I will evaluate its legitimacy based on how many childish sounding arguments I hear because obviously, they are not the real concern and people who defend them have zero credibility." Even among those who provide other arguments, it seems more common than not that they also assume that the fast takeoff scenario is less likely a priori because it's childish. Conversely, I've never heard anyone imply that fast takeoff must be true, or even likely, because it's exciting. It seems that, to have a similar effect, the excitement bias would have to do some heavy lifting in a purely subconscious way.

Given all this, the fact that I've seen far more discussion about the excitement-bias than the childishness-bias seems wrong.

Disclaimer: I don't think this is a strong argument that fast takeoff is likely, and I also don't think that bias towards excitement is weak in general – just that it's weak among the relevant class of people. Finally, I don't think this bias is common on LessWrong, only just about everywhere else.

How to Escape From Immoral Mazes

Previously in sequence and most on point: What is Success in an Immoral Maze?, How to Identify an Immoral Maze

This post deals with the goal of avoiding or escaping being trapped in an immoral maze, accepting that for now we are trapped in a society that contains powerful mazes.

We will not discuss methods of improving conditions (or preventing the worsening of conditions) within a maze, beyond a brief note on what a CEO might do. For a middle manager anything beyond not making the problem worse is exceedingly difficult. Even for the CEO this is an extraordinarily difficult task.

To rescue society as a whole requires collectively fighting back. We will consider such options in later posts.

For now, the problem statement is hard enough.

To reiterate the main personal-level takeaway

Being in a maze is not worth it. They couldn't pay you enough. Even if they could, they definitely don't. If you end up CEO, *you still lose*. These lives are not worth it. Do not be a middle manager at a major corporation or other organization that works like this. Do not sell your soul.

Problem statement

Increasingly, avoiding mazes is easier said than done.

First, one must identify them, for which the previous post offers a guide.

After that, there are still many hard problems to solve.

How do we avoid moral mazes? How do we justify that choice to others? What alternative choices do we have? What if we're already in a maze? What if we've already self-modified in ways that make it hard to extract ourselves? What if our human or social capital only pays off inside them?

What about if you are doing object-level work without anyone who reports to you, but you have a maze above you?

And for those who think this way, do I have a moral obligation to suffer and do this anyway, in order to maximize my charitable giving, or to otherwise do good?

How do we avoid immoral mazes?

Truly understand how painful it will be to interact with a maze even if you're not an employee. Know the signs, as discussed in the previous post. Keep a close eye out for mazes. Realize that you have other options. Choose other paths.

This isn't an 'all things being equal' choose other paths. This is making what *look* like major sacrifices and different life choices and profession choices, or taking big risks (that may or may not include starting a business or doing work outside of an organization) in order to have skin in the game. Really understand that the offer from even a relatively tolerable maze is much, much worse than it looks, and that opportunities outside mazes are often much better and more realistic than they look.

Young people starting out in the labor market often have The Fear that they will never find a job or never find a good job or another good job. If you are capable of getting this far, and you persevere, that is not true for you. A wide variety of jobs and other opportunities are out there.

I realize some people have already become so trapped in mazes that they cannot walk away.

If you actually can't walk away, see the last two questions.

What do you do if you find yourself inside a maze?

Quit. Seriously. Go do something else. Ideally, do it today.

At least start planning and looking. Every day there is another day you suffer, another day you invest your social and human capital in ways that can't be transferred, and another day you become infected by the maze that much more.

If you actually can't afford to quit, see the last three questions.

How do we justify our choice to others?

When I worked for a financial firm, the question 'what do you do?' (or, in a scarier form, 'who are you?' as implicitly defined by your job) had an easy answer. I work for (firm). One of the big benefits was being able to tell an easy, compact story of me and my life and work choices, that most found praiseworthy. It was easy. It was *comfortable.* It also worked wonders for things like renting an apartment or otherwise proving myself respectable.

A lot of the alternative answers that don't involve mazes give you a much better life and method of earning a living, but they do make answering the 'what do you do?' and 'how can I count on you to make rent or support a family?' questions trickier. One must acknowledge this.

It isn't only strangers you tell this story to. It is your friends. It is your family. It is also yourself.

I likely stayed at (firm) months longer than I should have due to being scared of not being able to tell this story anymore, especially to my wife and to myself, and having to instead tell a different one.

A lot of this fear is the expectation that others won't understand and won't accept our justifications. That does happen, but *far less than people typically expect or fear.* Most people are far more sympathetic than the inside view might suggest.

Even the internet is supportive. Which is not its style.

This is largely because, at least for now, there is a widespread cultural belief that one should do what you love, and be content in one's work. That work should provide meaning. That's not always a good rule or good idea, although it is a fine aspiration. Not everyone can have soul in the game. But almost everyone recognizes that it would be better if one did.

How do you go about telling your new story? (Justification continued)

Here's my take on how to approach this, based on my experience. Comments suggesting improvements or alternatives are highly encouraged.

There are two parts of this.

One is to figure out what you *are* doing, not only what you're *not* doing, and how to talk about that. Some of those answers are mostly culturally normal and comfortable, some of them are less so. Now that I can say 'I'm a game designer' that goes over quite well.

The most important things here are to make the thing you are doing sound *simple,* put it in terms that people can relate to, and to make it clear that you are comfortable and happy with it to the extent that this wouldn't be lying to people. If you're not comfortable and happy with it, people will pounce on that. It's also much better to be happy with what you are doing for your own sake, so that is something to work on, either looking to get to that place, or to finding another option where you can do that.

If you quit today without a plan, then what you will be doing is recovering from your experience and figuring out what to do next. I told that story for about a month. That story goes over better than you would expect – for a while. From a social (as well as financial) perspective you are most definitely on a clock. There are plenty of people who let that clock run out. But if it's two weeks in, own it.

The other part of your explanation is justifying why you're not doing the standard thing of indenturing yourself to a new maze, unless you have an obviously great alternative thing going. The worse your answer to part one sounds, the harder part two is going to be.

First try giving it to them straight. Tell them you find large corporations highly toxic and morally compromising. It left you a wreck. You have no interest in the lifestyle you would live or the person that you would be. If they are genuinely curious, you can point them to *Moral Mazes* itself or this series of posts, or explain further in your own words.

You can also use the culturally assigned incantations to explain your decision. Tell people you need to do what you are passionate about, to 'follow your heart/passion,' to do what you believe in, to 'help people' or 'make a difference.' To 'get your hands dirty' and 'do something real.' Some people appreciate wanting to 'be your own boss' or 'do it your way,' which are weaker, non-dystopian ways of sending the real message.

And of course, you can simply say 'it was making me miserable. I hated it. Doing this instead makes me happy.'

I've gotten into the habit of saying my job at (firm) was 'a poor fit' because I genuinely believe both that there were particular real needs they have that were expensive for me to provide, and that the firm was in many ways unusually great and unusually low on maze characteristics, and I do not think it is a mistake to take a job there if it would suit you. They treated me right and I don't want to throw anyone under the bus.

You *could* also, if you wanted to, use the question as an opportunity to do a public service and spread the word. That's supererogatory.

Another thing to keep in mind, as discussed in the next question, is that those pushing us towards mazes are often operating on traditions and heuristics that used to push towards virtuous action that led to happiness and real success. The world changed, and those traditions and heuristics started getting this wrong. This is highly sympathetic. It might help to approach from this perspective.

Most people get it. They don't *fully* get it unless they've been on in the inside and reflected upon what happened. But they do get that there's something soul-killing about working for the man and/or being lost in a maze of political actions.

Others won't get it.

What if my family or culture won't accept my justifications?

Some people won't get it. They will respond that all of this is excuses for not wanting to do hard work or make sacrifices. That this is how the 'real world' works. That it is 'time to grow up.' That a 'real adult/man/woman/hero/whomever' would suck it up and deal with it. That it is your responsibility to do so, for your family, for the world or for yourself. That 'get a steady job' is what good and responsible people do. That this is how one survives in today's world, and how one gets to raise and support a family.

Often people who are counting on you, usually family members, will effectively let you know that while they care a non-zero amount about whether your life experience is miserable, or what impact your work has on the world, or what upside or opportunities for personal growth you might have, what they *actually* care about is whether you are projecting the illusion of security. They want to mentally cache that you/they 'are going to be OK' and that 'everything is all right.'

This has remarkably little to do with actual security. Jobs in many mazes are not especially secure. Others are secure barring disruption of the underlying order, if you are willing to pay the prices discussed, and tie all your human and social capital to the maze.

The security they seek is the security of the banker who loses money when everyone around him loses money. This is useful to the banker because one cannot then scapegoat and fire him if he has bad luck and does poorly. This is useful to those in a maze and those who tie their fates to them, because they hope they will similarly seem responsible and legitimate, and thus worthy of sympathy and assistance if things go poorly.

It is a self-reinforcing charade. People demand this illusion of legitimacy to protect against others' accusations of illegitimacy. They will defend this charade even if times change, the mazes mostly fail and those who are doing real things succeed, and attempt to forcibly transfer wealth from people doing real things to those who previously worked in mazes. All of that doesn't make participation in the charade

worthless. You can even hope to benefit from the expropriations. But it is important to know that it is a charade.

If you face a family and/or culture that demands devotion of one's entire life to the illusion of respectability, have sympathy for those making these demands. In the past, when these traditions and heuristics developed, the illusion of respectability corresponded to real work and other worthy virtues that lead to happiness and true success. If they fail to realize the change, and update accordingly, that is at least somewhat understandable.

If attempts to make them realize this change or accept your perspective fail, one must treat them the same way one treats anything else that is out to get you. If their demands cannot be satisfied in a way you can accept, because they will simply demand more until it is something you cannot accept, then attempting to satisfy their demands is folly.

If you previously chose those around you based upon being a member of a maze, then it is plausible that having invested in those people and relationships it makes sense to stick around. It is also plausible that being around them afterwards no longer makes sense.

I know it is easy to say and tough to act upon, but hopefully in time either they will understand and/or come around, or you will realize that life is better without them.

What if you've already self-modified too much?

This sucks quite a lot.

After a while, those little status differences and little battles start to deeply matter. Other things matter less and less. Humans can adapt to many things. Giving up all that likely fills you with deep existential dread.

Even worse, you've sculpted everything else, including your friends and often your family, around these obsessions. You, and often they, depend on the currently steady money *and the illusion of security* the maze provides. Without that, things could rapidly fall apart.

The good news is, you've figured out that this happened. Perhaps you haven't self-modified as much as you've feared, or you have a path back to undoing this. Often the modifications start reverting once you extract yourself from the situation. Often you're deeply miserable, in a way that those around you at least subconsciously know quite well. Those around you often realize this long before the person realizes it about themselves. If you tell the people you care about what's really going on, if they're worth keeping, they'll almost always be supportive.

I've seen a number of people realize they hated their jobs and needed to go. Almost all of them got lots of support when they got the courage to say it out loud.

A note I got on a previous draft: "The happiest Uber drivers I have seen used to be middle management."

If anything, I see many people around me being *too supportive* of opting out of working, or in a sense out of life, entirely. It is important to help and encourage people to do more things.

See the question above on how to explain your choices and situation to others.

So my first suggestion is to admit to yourself what is happening to you. Take an inventory. Concretely observe *what is actually going on around you,* without excuses or euphemisms, and what that is doing to your brain and your life.

Then *tell the people who care about you.* And go from there.

Are mazes are where our human and/or social capital pays off?

Note that when I say ‘pays off’ here, I mean *maximally* pays off. If you have the skills and opportunity to advance inside a maze, you have the skills and opportunity to take a lower level position at a smaller institution, and still earn a reasonable living. That does not mean that this transition would not be painful, or that you could maintain your current lifestyle, or that your family and friends would stand by you, *but you definitely have that option.*

If you are an academic with a PhD, and notice your academic institution is a maze, keep in mind that the entire concern about you only getting paid off in academia probably *simply isn’t true*. Academics typically get substantial pay bumps when they move to private industry.

A lot of other professionals are similarly buying the security and familiarity of what they’re used to, rather than being paid off with dollars.

One must still be careful to ensure not to jump from one maze into another.

Then there’s big corporations. Big corporations *really do* pay better than smaller businesses if you can’t get equity in either business. The recent history of domestic ‘income inequality’ is in large part inequality *between firms* as bigger and more successful firms pay higher salaries even to their lower tiers, and have more higher tiers in which to earn yet more.

There are three theories I know about for why big corporations pay more.

Theory one is that big corporations are more likely to have O-ring production functions or otherwise benefit more from higher quality workers, so they pay more in order to attract better workers.

Theory two is that big corporations make more money per employee, and are big enough to potentially support unions, so employees demand and receive more of that pay.

Theory three is that working in a big corporation sucks, and employees realize this to at least some extent, so employees demand more money in order to be willing to work there.

If working at a major corporation is a major life cost, and working in management a bigger one, and these come with higher pay, than a lot of income inequality in developed countries does not represent a gap in desired life outcomes, and it might be *more unfair* if that part of the gap was closed.

A lot more of that pay gap is that some professions engage in rent seeking behavior to extract resources. Some big examples are finance, law, education/academia, and medicine. Again, that comes with much better pay.

It also usually comes with a big time investment in the development of the relevant social capital, human capital and credentials you need to succeed. If you went to medical school or law school or worked hard to get a tenure track, and later realize that your profession is a maze (I'm making no claim here that these professions are or aren't mazes in general or how intense those mazes might be, although some central organizations within them clearly are very intense mazes), walking away from that is going to be expensive.

This is doubly true for human capital *within a single organization.* When I took a job at a financial firm, they spent a large part of my first *two years* training me. The first year had a lot of firm-specific detail but was mostly training about markets and trading in general, that applies everywhere there are markets. It was fascinating, and pretty great. Five stars, would study again, especially given they paid me rather than the other way around. The second year still had a lot of training and learning, but increasingly it was about the specific problems I was working on, developing relationships with and learning about coworkers and organizational structures and how we did things, and other information specific to the firm. This was less fun, and when I left, it became worthless.

I had another job I stayed in for five years. This was also a place I got to observe transition from mostly not being a maze into becoming one over time, although that's a story I can't tell online.

Early on there was a lot of learning, a lot of which was very specific to our business, but a lot of which applies universally. I worked mainly with one particular person, who knew what we were doing and cared about us doing it well. It provided great experience.

By the third year, I was learning about our specific products and customers and dynamics, in increasingly arcane fashion. I was also forced to interact increasingly with the maze growing around us, spending more time making bosses and others like what they saw rather than doing what was right for the business. I was unable to get the resources to enhance our performance, despite yearly returns on investment obviously well above 100%. I made an effort to switch over into problems that were both more valuable and offered more room for growth, both for me and for the business, and which I could tackle with the resources available.

By the fifth year, I wasn't developing any skills that would be useful elsewhere, except that I was now learning to code because I got tired of no one being able to code what I needed. This brought me from 'can code but not in an actually useful way' to 'can code real things that are useful, but badly/slowly.'

Note that the managers in Moral Mazes who succeed were always moving around to bigger and better things. If they weren't, they instead moved on to similar and different thus hard to compare things to preserve the illusion of career momentum. If you have adopted the maze nature, many of the skills you have learned doing so translate to other mazes. Your existing within-maze status can often also be transferred to your new location, but only if you continue to be seen as a winner. If you're a loser, no one else will want you, and moving on will mean moving down.

That means that once your path in the maze is stalled, even though you have invested a lot to get to this point, recovering your momentum is going to be extremely difficult. If you are not satisfied with your current role, your human capital is a lot less valuable than it naively appears to be, because it no longer has much upside even on its own terms. The fall from where you are to starting over can still be large.

The best feature of an academic maze is that they have a perfectly designed system in which to not care about getting ahead, which The Gervais Principle calls a loser, and which academia calls tenure.

The pattern remains. The more you dedicate time to a path, both a profession and a particular job, the more you give up when you leave and the less of your time you can carry with you. Many people don't have great options. The job market isn't that great out there if you don't want to be coding and don't have an in with the rent seekers, and can't use the skills you've developed, or do the thing that legibly follows from your resume.

If mazes are where my social and/or human capital pays off, what should I do?

Let us ignore here *why* your capital pays off best in a maze. It does not much matter to your decision, in important senses, to what extent rent seeking, theft, coercion, fraud or even systems designed explicitly to make your skills not transfer to honest work are or aren't responsible for this being the case. For whatever reason, often events conspire to prevent you from efficiently plying your trade (or in some cases, plying it at all), where you hold comparative advantage, without being part of a maze.

Some people really do have a dilemma, where they can either do something menial and mindless that still gets them abused and doesn't pay much, if they can find work at all. Or they can go out on a limb that looks super risky and likely to fail, and/or that requires years without compensation. Or they can keep working in the maze.

It is important that *vastly more people **think** they are in this position, than **are** in this position.* If you think you are in this position, consider the possibility that you are mistaken. Consider all the alternatives. Consider how much the reduction in medium-term funds and superficial status would actually matter to you. Consider how much of what's holding you back is simply The Fear in some form.

Imagine exactly how relieved you'd be to be out of there. Remember that *even if leaving really is super painful,* involving a large reduction in consumption levels and superficial status and standard of living, and the abandonment of large sunk costs, that doesn't mean it isn't Worth It.

My first line of response to this dilemma is exactly what you would expect: *Consider leaving anyway.* But I admit that isn't *always* the right answer. In some cases, things really have gone too far, you have too many promises to keep and too many sunk costs.

Become a Loser

The next line of defense is to *become a loser,* in the sense laid out in The Gervais Principle. A loser does not strive to get ahead while at work. A loser finds their value in

other places than work. At work, they pride themselves on putting forward *at most* the minimum amount of effort to get the job done.

The Gervais Principle can be seen as the prequel to *Moral Mazes,* dealing with life at lower levels of mazes that have to interact with the real world. Mazes need, as several quotes describe, people who keep their heads down and 'do their job' with no ambitions for further advancement. Ideally one does this as low on the totem pole as one can stomach and afford, as the life that results is far less odious and taxing.

By declaring themselves as neutral and not a threat, such people are often left mostly alone if they're important to the system continuing to run. They can now reclaim some slack and a personal life. It's not a great solution. You're still holding up the maze. You're still interacting with it. You'll still have to make severe moral sacrifices. But to some extent, some of the time, you can pick and choose what to have no part in.

Over time, your position likely will slowly degrade. Eventually this may lead you to leave. Hopefully by then you'll have been able to save enough and be prepared enough to be ready for that. If you're stuck in a maze, the least you can do is turn a healthy monthly profit.

Take Risks

The final line of defense I can come up with is to take big bold risks. Either stand up for what you believe in or gamble to advance your own situation. Sometimes this will work, your situation will improve and you'll learn your situation was better than you thought. Other times they'll backfire, and you'll learn your situation was worse than you thought and is now worse than that. Remember that if you get fired from a job you don't want, that can be a big win, because you might not have had the courage to leave on your own and you might even get severance and unemployment.

The real danger is often not that you get fired. It's that you become 'dead without knowing it' as in this quote:

> You can put the damper on anyone who works for you very easily and that's why there's too much chemistry in the corporation. There's not enough objective information about people. When you really want to do somebody in, you just say, well, he can't get along with people. That's a big one. And we do that constantly. What that means, by the way, is that he pissed me off; he gave evidence of his frustration with some situation. Another big one is that he can't manage—he doesn't delegate or he doesn't make his subordinates keep his commitments. So in this sort of way, a consensus does build up about a person and a guy can be dead and not even know it. (Location 1475, Quote 10)

This can lead you to waste years of your life struggling for something you had no chance of getting. This is one reason why a great way to take risk is to force the issue, asking for or demanding raises or promotions. It avoids this danger. The more uncertain you are about where you stand, the more you should take risk to create clarity.

At all my jobs in mazes, I would have greatly benefited from taking greater risks to create clarity, *regardless of the outcome.*

Can You Change Things From the Top?

If you by some miracle reach the top with your soul intact, now you can try and change the system. Or at least you can do harm reduction in earnest. One shouldn't give this much hope or weight, since such intentions rarely survive that long, and doing anything lasting about it will still be quite hard. I don't know what would work.

My friends and I have talked to several people who have reached the top. Many of them understand what the process has done to them, but don't know how to fix themselves or the system. It isn't cheap or easy to reverse or even halt the damage.

It is unlikely you can have much impact without reaching the actual top and becoming CEO. If you do become CEO, you may have a short window in which you can 'clean house' the way that maze CEOs do, and put people opposed to mazes into key positions where they can clean their houses in turn. You can then combine that with other efforts, and maybe get somewhere, but I don't have the insights necessary to say much more, and such efforts will be exceedingly difficult. The maze will fight back.

I strongly believe it is much easier to build a new system from scratch than to 'change the system from within.'

What about if you are doing object-level work without anyone who reports to you, but you have a maze above you?

In *Moral Mazes* such workers are said to be 'on the line.'

Details of this situation will determine to what extent this represents being stuck in the maze, versus to what extent this represents doing regular object-level work.

What are you actually doing all day? What are your incentives?

If your essential scenario is *given an object-level job to do and do it,* that is mostly fine.

If your essential scenario is not that, it is less fine, but it is still far better than being a middle manager. It's not *good* to have a bull**** job, but it's not the nightmare we're describing elsewhere.

Consider the car salesmen from Imperfect Competition.

One can imagine a car dealership as no different from the local hardware store, buying useful tools wholesale and selling them at a higher price to customers who want to buy and use those tools, and the only difference is that you sell 0.1% as many tools for a thousand times the price. One can also imagine that the demands of the car corporation, and the incentives they provide, and the misinformation they spread, and the regulations they twist and engineer, and so forth, end up with you effectively stuck in the maze. The truth is presumably somewhere in between – you see insane things around quotas and regulations and advertising campaigns you cannot control, and the dealerships have their own issues of their own design, but you are still mostly working for a small actual business most of the time.

The same would go for the workers in The Office, as analyzed in The Gervais Principle. Michael is largely in maze hell. Jim spends a lot of time avoiding maze hell. Most of the workers have to deal with the craziness and what it does to the business, but this is only ordinary soul crushing and not what middle managers deal with.

Jim's situation on The Office is the biggest problem. There is no future. The only way up, to better your work situation, would be to dive into the maze. If you do that while not buying into the system, it will go badly for you on all levels. If you do buy in, then you've fully made the big mistake I'm warning against.

Consider the Uber drivers, some of whom are reported to be happy refugees from middle management.

To the extent that the driver is offered rides, chooses to accept them individually, and gets paid for each ride provided, the driver is good. They set their hours and level of effort. There is word that Uber and its ilk are now using algorithmic systems and various overall incentives to try and ensnare their drivers more broadly into the system, which would be worse, but the core experience is still one of real work.

Consider a software engineer, given specific tasks to code and coding them. That seems likely to be mostly fine.

Consider a worker that is literally 'on the line' in a manufacturing plant that makes physical objects. It is not the best or most compensated work, but you are mostly free from the maze.

Being 'on the line' and continuing to do real work is miles behind doing real work where you have skin in the game, but if you get to dodge the worst of all this, it is a reasonable temporary fallback if you lack alternatives. Look carefully at details.

If you are a manager but not a middle manager (e.g. no one who reports to you has anyone who reports to them), and the group of people you manage has object-level tasks to do together, you aren't automatically doomed, but there is great danger lurking, including the risk you will be promoted.

Do I have a moral obligation to work in mazes to maximize my charitable giving?

No. You don't.

This post has done its best to deliberately ignore the moral costs of participating in mazes, *because avoiding them is already over-determined without that.*

I want to make it clear that I'm *not relying* on moral concerns.

But if that's what you are concerned about, moral concerns work in the opposite direction. Making the world more and more maze-like by embracing the system, and engaging in zero-sum competitions to extract resources, while making your life miserable, is the opposite of a moral obligation.

It may help to remember that a drowning child is hard to find.

Moral systems that imply that subjecting oneself to torture in the service of immoral mazes or other harmful systems, for the purposes of allowing other such systems to then extract those resources from you, is a moral obligation, are not likely to be good ideas, or to have your or humanity's best interests at heart.

Subscripting Typographic Convention For Citations/Dates/Sources/Evidentials: A Proposal

Reviving an old General Semantics proposal: borrowing from scientific notation and using subscripts like 'Gwern$_{2020}$' for denoting sources (like citation, timing, or medium) might be a useful trick for clearer writing, compared to omitting such information or using standard cumbersome circumlocutions.

Moved to gwern.net

The two-layer model of human values, and problems with synthesizing preferences

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

I have been thinking about Stuart Armstrong's preference synthesis research agenda, and have long had the feeling that there's something off about the way that it is currently framed. In the post I try to describe why. I start by describing my current model of human values, how I interpret Stuart's implicit assumptions to conflict with it, and then talk about my confusion with regard to reconciling the two views.

The two-layer/ULM model of human values

In Player vs. Character: A Two-Level Model of Ethics, Sarah Constantin describes a model where the mind is divided, in game terms, into a "player" and a "character". The character is everything that we consciously experience, but our conscious experiences are not our true reasons for acting. As Sarah puts it:

> In many games, such as Magic: The Gathering, Hearthstone, or Dungeons and Dragons, there's a two-phase process. First, the player constructs a *deck* or *character* from a very large sample space of possibilities. This is a particular combination of strengths and weaknesses and capabilities for action, which the player thinks can be successful against other decks/characters or at winning in the game universe. The choice of deck or character often determines the strategies that deck or character can use in the second phase, which is actual gameplay. In gameplay, the character (or deck) can only use the affordances that it's been previously set up with. This means that there are two separate places where a player needs to get things right: first, in designing a strong character/deck, and second, in executing the optimal strategies for that character/deck during gameplay. [...]
>
> The idea is that human behavior works very much like a two-level game. [...] The player determines what we find rewarding or unrewarding. The player determines what we notice and what we overlook; things come to our attention if it suits the player's strategy, and not otherwise. The player gives us emotions when it's strategic to do so. The player sets up our subconscious evaluations of what is good for us and bad for us, which we experience as "liking" or "disliking."
>
> The character is what *executing the player's strategies feels like from the inside*. If the player has decided that a task is unimportant, the character will experience "forgetting" to do it. If the player has decided that alliance with someone will be in our interests, the character will experience "liking" that person. Sometimes the player will notice and seize opportunities in a very strategic way that feels to the character like "being lucky" or "being in the right place at the right time."

> This is where confusion often sets in. People will often protest “but I *did* care about that thing, I just forgot” or “but I’m *not* that Machiavellian, I’m just doing what comes naturally.” This is true, because when we talk about ourselves and our experiences, we’re speaking “in character”, as our character. The strategy is not going on at a conscious level. In fact, I don’t believe we (characters) have direct access to the player; we can only *infer* what it’s doing, based on what patterns of behavior (or thought or emotion or perception) we observe in ourselves and others.

I think that this model is basically correct, and that our emotional responses, preferences, etc. are all the result of a deeper-level optimization process. This optimization process, then, is something like that described in The Brain as a Universal Learning Machine:

> The universal learning hypothesis proposes that *all* significant mental algorithms are learned; nothing is innate except for the learning and reward machinery itself (which is somewhat complicated, involving a number of systems and mechanisms), the initial rough architecture (equivalent to a prior over mindspace), and a small library of simple innate circuits (analogous to the operating system layer in a computer). In this view the mind (software) is distinct from the brain (hardware). The mind is a complex software system built out of a general learning mechanism. [...]
>
> An initial untrained seed ULM can be defined by 1.) a prior over the space of models (or equivalently, programs), 2.) an initial utility function, and 3.) the universal learning machinery/algorithm. The machine is a real-time system that processes an input sensory/observation stream and produces an output motor/action stream to control the external world using a learned internal program that is the result of continuous self-optimization. [...]
>
> The key defining characteristic of a ULM is that it uses its universal learning algorithm for continuous recursive self-improvement with regards to the utility function (reward system). We can view this as second (and higher) order optimization: the ULM optimizes the external world (first order), and also optimizes its own internal optimization process (second order), and so on. Without loss of generality, any system capable of computing a large number of decision variables can also compute internal self-modification decisions.
>
> Conceptually the learning machinery computes a probability distribution over program-space that is proportional to the expected utility distribution. At each timestep it receives a new sensory observation and expends some amount of computational energy to infer an updated (approximate) posterior distribution over its internal program-space: an approximate 'Bayesian' self-improvement.

Rephrasing these posts in terms of each other, in a person's brain "the player" is the underlying learning machinery, which is searching the space of programs (brains) in order to find a suitable configuration; the "character" is whatever set of emotional responses, aesthetics, identities, and so forth the learning program has currently hit upon.

Many of the things about the character that seem fixed, can in fact be modified by the learning machinery. One's sense of aesthetics can be updated by propagating new facts into it, and strongly-held identities (such as "I am a technical person") can change in response to new kinds of strategies becoming viable. Unlocking the

Emotional Brain describes a number of such updates, such as - in these terms - the ULM eliminating subprograms blocking confidence after receiving an update saying that the consequences of expressing confidence will not be as bad as previously predicted.

Another example of this kind of a thing was the framework that I sketched in Building up to an Internal Family Systems model: if a system has certain kinds of bad experiences, it makes sense for it to spawn subsystems dedicated to ensuring that those experiences do not repeat. Moral psychology's social intuitionist model claims that people often have an existing conviction that certain actions or outcomes are bad, and that they then level seemingly rational arguments for the sake of preventing those outcomes. Even if you rebut the arguments, the conviction remains. This kind of a model is compatible with an IFS/ULM style model, where the learning machinery sets the goal of preventing particular outcomes, and then applies the "reasoning module" for that purpose.

Qiaochu Yuan notes that once you see people being upset at their coworker for criticizing them and you do therapy approaches with them, and this gets to the point where they are crying about how their father never told them that they were proud of them... then it gets really hard to take people's reactions to things at face value. Many of our consciously experienced motivations, actually have nothing to do with our real motivations. (See also: Nobody does the thing that they are supposedly doing, The Elephant in the Brain, The Intelligent Social Web.)

Preference synthesis as a character-level model

While I like a lot of the work that Stuart Armstrong has done on synthesizing human preferences, I have a serious concern about it which is best described as: everything in it is based on the character level, rather than the player/ULM level.

For example, in "Our values are underdefined, changeable, and manipulable", Stuart - in my view, correctly - argues for the claim stated in the title... except that, it is not clear to me to what extent the things we intuitively consider our "values", are actually our values. Stuart opens with this example:

> When asked whether "communist" journalists could report freely from the USA, only 36% of 1950 Americans agreed. A follow up question about Amerian journalists reporting freely from the USSR got 66% agreement. When the order of the questions was reversed, 90% were in favour of American journalists - and an astounding 73% in favour of the communist ones.

From this, Stuart suggests that people's values on these questions should be thought of as *underdetermined*. I think that this has a grain of truth to it, but that calling these opinions "values" in the first place is misleading.

My preferred framing would rather be that people's *values* - in the sense of some deeper set of rewards which the underlying machinery is optimizing for - are in fact underdetermined, *but* that is not what's going on in this particular example. The order of the questions does not change those values, which remain stable under this kind of a consideration. Rather, consciously-held political opinions are *strategies* for carrying out the underlying values. Receiving the questions in a different order caused the

system to consider different kinds of information when it was choosing its initial strategy, causing different strategic choices.

Stuart's research agenda does talk about incorporating meta-preferences, but as far as I can tell, all the meta-preferences are about the character level too. Stuart mentions "I want to be more generous" and "I want to have consistent preferences" as examples of meta-preferences; in actuality, these meta-preferences might exist because of something like "the learning system has identified generosity as a socially admirable strategy and predicts that to lead to better social outcomes" and "the learning system has formulated consistency as a generally valuable heuristic and one which affirms the 'logical thinker' identity, which in turn is being optimized because of its predicted social outcomes".

My confusion about a better theory of values

If a "purely character-level" model of human values is wrong, how do we incorporate the player level?

I'm not sure and am mostly confused about it, so I will just babble & boggle at my confusion for a while, in the hopes that it would help.

The optimistic take would be that there exists some set of universal human values which the learning machinery is optimizing for. There exist various therapy frameworks which claim to have found something like this.

For example, the NEDERA model claims that there exist nine negative core feelings whose avoidance humans are optimizing for: people may feel Alone, Bad, Helpless, Hopeless, Inadequate, Insignificant, Lost/Disoriented, Lost/Empty, and Worthless. And pjeby mentions that in his empirical work, he has found three clusters of underlying fears which seem similar to these nine:

> For example, working with people on self-image problems, I've found that there appear to be only three critical "flavors" of self-judgment that create life-long low self-esteem in some area, and associated compulsive or avoidant behaviors:
>
> Belief that one is bad, defective, or malicious (i.e. lacking in care/altruism for friends or family)
>
> Belief that one is foolish, incapable, incompetent, unworthy, etc. (i.e. lacking in ability to learn/improve/perform)
>
> Belief that one is selfish, irresponsible, careless, etc. (i.e. not respecting what the family or community values or believes important)
>
> (Notice that these are things that, if you were bad enough at them in the ancestral environment, or if people only **thought** you were, you would lose reproductive opportunities and/or your life due to ostracism. So it's reasonable to assume that we have wiring biased to treat these as high-priority *long-term* drivers of compensatory signaling behavior.)

> Anyway, when somebody gets taught that some behavior (e.g. showing off, not working hard, forgetting things) *equates* to one of these morality-like judgments as a *persistent quality* of themselves, they often develop a compulsive need to prove otherwise, which makes them choose their goals, not based on the goal's actual utility to themself or others, but rather based on the goal's perceived value as a means of virtue-signalling. (Which then leads to a pattern of continually trying to achieve similar goals and either failing, or feeling as though the goal was unsatisfactory despite succeeding at it.)

So - assuming for the sake of argument that these findings are correct - one might think something like "okay, here are the things the brain is trying to avoid, we can take those as the basic human values".

But not so fast. After all, emotions are all computed in the brain, so "avoidance of these emotions" can't be the only goal any more than "optimizing happiness" can. It would only lead to wireheading.

Furthermore, it seems like one of the things that the underlying machinery *also* learns, is *situations in which it should trigger these feelings*. E.g. feelings of irresponsibility can be used as an internal carrot and stick scheme, in which the system comes to predict that if it will feel persistently bad, this will cause parts of it to pursue specific goals in an attempt to make those negative feelings go away.

Also, we are not *only* trying to avoid negative feelings. Empirically, it doesn't look like happy people end up doing *less* than unhappy people, and guilt-free people may in fact do more than guilt-driven people. The relationship is nowhere linear, but it seems like there are plenty of happy, energetic people who are happy *in part because* they are doing all kinds of fulfilling things.

So maybe we could look at the inverse of negative feelings: positive feelings. The current mainstream model of human motivation and basic needs is self-determination theory, which explicitly holds that there exist three separate basic needs:

> **Autonomy:** people have a need to feel that they are the masters of their own destiny and that they have at least some control over their lives; most importantly, people have a need to feel that they are in control of their own behavior.
>
> **Competence:** another need concerns our achievements, knowledge, and skills; people have a need to build their competence and develop mastery over tasks that are important to them.
>
> **Relatedness (also called Connection):** people need to have a sense of belonging and connectedness with others; each of us needs other people to some degree

So one model could be that the basic learning machinery is, first, optimizing for avoiding bad feelings; and then, optimizing for things that have been associated with good feelings (even when doing those things is locally unrewarding, e.g. taking care of your children even when it's unpleasant). But this too risks running into the wireheading issue.

A problem here is that while it might make intuitive sense to say "okay, if the character's values aren't the real values, let's use the player's values instead", the split isn't actually anywhere that clean. In a sense the player's values are the real

ones - but there's also a sense in which the player doesn't *have* anything that we could call values. It's just a learning system which observes a stream of rewards and optimizes it according to some set of mechanisms, and even the reward and optimization mechanisms themselves may end up getting at least partially rewritten. The underlying machinery has no idea about things like "existential risk" or "avoiding wireheading" or necessarily even "personal survival" - thinking about those is a character-level strategy, even if it is chosen by the player using criteria that it does not actually understand.

For a moment it felt like looking at the player level would help with the underdefinability and mutability of values, but the player's values seem like they could be *even less* defined and *even more* mutable. It's not clear to me that we can call *them* values in the first place, either - any more than it makes meaningful sense to say that a neuron in the brain "values" firing and releasing neurotransmitters. The player is just a set of code, or going one abstraction level down, just a bunch of cells.

To the extent that there exists something that intuitively resembles what we call "human values", it feels like it exists in some hybrid level which incorporates parts of the player *and* parts of the character. That is, assuming that the two can even be very clearly distinguished from each other in the first place.

Or something. I'm confused.

Disasters

If there were a natural disaster tomorrow and it took about two weeks to get things working again, how many people would be ok for food, water, and other necessities? I'm guessing below 5%, but I think this level of preparedness would be a good goal for most people who can afford it. Why don't people plan for potential disasters? Some possibilities:

- They don't think disasters are likely. On the other hand, I also don't think disasters are likely! While we have extra water in the basement, I think the chances we'll need it sometime during my life are only maybe 2%. Since it's not expensive, and if we do need it we'll be incredibly happy to have it, I think it's worth setting up.

  It does matter a lot whether the chances are ~2% or 0.0002%, but if you think your lifetime chance of being impacted by a serious disaster is under 1% I'd encourage you to think about historical natural disasters in your area (earthquakes, floods, hurricanes, wildfires, etc) plus the risk of potential human-caused disasters (nuclear war, epidemics, civil war, economic collapse, etc).

- It's weird. Most people don't do it, and a heuristic of "do the things other people do" is normally a pretty good one. In this case, though, I think we should be trying to change what's normal. The government agrees; the official recommendations involve a lot more preparation than people typically do.

- They can't afford the money, time, or thought. Many people are in situations where planning for what's likely to happen in the next couple months is hard enough, let alone for things that have a low single digits chance of happening ever. This can't explain all of it, though, because even people who do have more time and money also haven't generally thought through simpler preparations.

- They don't think preparation is likely to be useful. If there's a nuclear strike we're all dead anyway, right? Except most disasters, even nuclear ones, aren't this binary. Avoiding exposure to radiation and having KI available can help your long-term chances a lot. Many disasters (nuclear, earthquake, epidemic, severe storm) are ones where having sufficient supplies to stay at home for weeks would be very helpful. If you think preparation wouldn't help and you haven't, say, read through the suggestions on ready.gov, I'd recommend doing that.

- They're used to local emergencies. We generally have a lot more experience with things like seeing houses burn down, knowing people who've become unable to work, or having family members get very sick. These can be major problems on a personal scale, but families, society, government, and infrastructure will generally still be intact. We can have insurance and expect that it will pay out; others in our families and communities may be able to help us. Things that affect a few people in a region or community at a time are the sort of things societies have the spare capacity for and figure out how to handle. A regional disaster works very differently, and makes planning in advance much more worthwhile.

- They expect to see it coming. Forecasting is good enough that we're very unlikely to be surprised by a hurricane, but for now an earthquake could still come out of nowhere. Others seem like the kind of thing we ought to be able to

anticipate, but are tricky: it's hard to see an economic collapse coming because economic confidence is anti-inductive and we tend to suddenly go from "things are good" to "things are very much not good". Paying attention is valuable, but it's not sufficient.

- They're not considering how bad things can be. For many of us our daily experience is really very good: high quality plentiful food and drink, comfortable and sufficient clothing, interesting things to do, good medical care. When you consider how bad a disaster can be, things that would improve your life a lot in very rare circumstances can make a lot of sense.

- They're not sure what to do. This is pretty reasonable: there's a ton of writing, often aimed at people who've gotten really into prepping, and not much in the way of "here are a few things to do if you want to allocate a weekend morning to getting into a better place". Storing extra water (~15gal/person), food (buy extra non-perishables and rotate through them), and daily medications, however, goes a long way. For a longer list, this guide seems pretty good. (Though they're funded by affiliate links so they have incentives to push you in the "buying things" direction.)

None of these seem very compelling to me, aside from cost, and the cost of basic preparations is pretty low. I think most people who can afford to would benefit a lot in expectation to put some time into thinking through

[AN #83]: Sample-efficient deep learning with ReMixMatch

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Highlights

ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring *(David Berthelot et al)* (summarized by Dan H): A common criticism of deep learning is that it requires far too much training data. Some view this as a fundamental flaw that suggests we need a new approach. However, considerable data efficiency is possible with a new technique called ReMixMatch. ReMixMatch on CIFAR-10 obtains 84.92% accuracy using only 4 labeled examples per class. Using 250 labeled examples, or around 25 labeled examples per class, a ReMixMatch model on CIFAR-10 has 93.73% accuracy. This is approximately how well a vanilla ResNet does on CIFAR-10 with 50000 labeled examples. Two years ago, special techniques utilizing 250 CIFAR-10 labeled examples could enable an accuracy of approximately 53%. ReMixMatch builds on MixMatch and has several seemingly arbitrary design decisions, so I will refrain from describing its design. In short, deep networks do not necessarily require large labeled datasets.

And just yesterday, after this summary was first written, the FixMatch paper got even better results.

Previous newsletters

In last week's email, two of Flo's opinions were somehow scrambled together. See below for what they were supposed to be.

Defining and Unpacking Transformative AI *(Ross Gruetzemacher et al)* (summarized by Flo): Focusing on the impacts on society instead of specific features of AI systems makes sense and I do believe that the shape of RTAI as well as the risks it poses will depend on the way we handle TAI at various levels. More precise terminology can also help to prevent misunderstandings, for example between people forecasting AI and decision makers.

When Goodharting is optimal: linear vs diminishing returns, unlikely vs likely, and other factors *(Stuart Armstrong)* (summarized by Flo): I enjoyed this article and the proposed factors match my intuitions. There seem to be two types of problems: extreme beliefs and concave Pareto boundaries. Dealing with the second is more important since a concave Pareto boundary favours extreme policies, even for moderate beliefs. Luckily, diminishing returns can be used to bend the Pareto

boundary. However, I expect it to be hard to find the correct rate of diminishing returns, especially in novel situations.

Technical AI alignment

Iterated amplification

AI Safety Debate and Its Applications *(Vojta Kovarik)* (summarized by Rohin): This post defines the components of a debate (AN #5) game, lists some of its applications, and defines truth-seeking as the property that we want. Assuming that the agent chooses randomly from the possible Nash equilibria, the truth-promoting likelihood is the probability that the agent picks the actually correct answer. The post then shows the results of experiments on MNIST and Fashion MNIST, seeing comparable results to the original paper.

(When) is Truth-telling Favored in AI debate? *(Vojtěch Kovařík et al)* (summarized by Rohin): Debate (AN #5) aims to train an AI system using self-play to win "debates" which aim to convincingly answer a question, as evaluated by a human judge. The main hope is that the equilibrium behavior of this game is for the AI systems to provide true, useful information. This paper studies this in a simple theoretical setting called *feature debates*. In this environment, a "world" is sampled from some distribution, and the agents (who have perfect information) are allowed to make claims about real-valued "features" of the world, in order to answer some question about the features of the world. The judge is allowed to check the value of a single feature before declaring a winner, but otherwise knows nothing about the world.

If either agent lies about the value of a feature, the other agent can point this out, which the judge can then check; so at the very least the agents are incentivized to honestly report the values of features. However, does this mean that they will try to answer the full question truthfully? If the debate has more rounds than there are features, then it certainly does: either agent can unilaterally reveal every feature, which uniquely determines the answer to the question. However, shorter debates need not lead to truthful answers. For example, if the question is whether the first K features are all 1, then if the debate length is shorter than K, there is no way for an agent to prove that the first K features are all 1.

Rohin's opinion: While it is interesting to see what doesn't work with feature debates, I see two problems that make it hard to generalize these results to regular debate. First, I see debate as being truth-seeking in the sense that the answer you arrive at is (in expectation) more accurate than the answer the judge would have arrived at by themselves. However, this paper wants the answers to actually be *correct*. Thus, they claim that for sufficiently complicated questions, since the debate can't reach the right answer, the debate isn't truth-seeking -- but in these cases, the answer is still in expectation more accurate than the answer the judge would come up with by themselves.

Second, feature debate doesn't allow for decomposition of the question during the debate, and doesn't allow the agents to challenge each other on particular questions. I think this limits the "expressive power" of feature debate to P, while regular debate reaches PSPACE, and is thus able to do much more than feature debate. See this comment for more details.

Read more: Paper: (When) Is Truth-telling Favored in AI Debate?

Mesa optimization

Malign generalization without internal search *(Matthew Barnett)* (summarized by Rohin): This post argues that agents can have capability generalization without objective generalization (AN #66), *without* having an agent that does internal search in pursuit of a simple mesa objective. Consider an agent that learns different heuristics for different situations which it selects from using a switch statement. For example, in lunar lander, if at training time the landing pad is always red, the agent may learn a heuristic about which thrusters to apply based on the position of red ground relative to the lander. The post argues that this selection across heuristics could still happen with very complex agents (though the heuristics themselves may involve search).

Rohin's opinion: I generally agree that you could get powerful agents that nonetheless are "following heuristics" rather than "doing search"; however, others with differing intuitions did not find this post convincing.

Agent foundations

Embedded Agency via Abstraction *(John S Wentworth)* (summarized by Asya): Embedded agency problems (AN #31) are a class of theoretical problems that arise as soon as an agent is part of the environment it is interacting with and modeling, rather than having a clearly-defined and separated relationship. This post makes the argument that before we can solve embedded agency problems, we first need to develop a theory of *abstraction*. *Abstraction* refers to the problem of throwing out some information about a system while still being able to make predictions about it. This problem can also be referred to as the problem of constructing a map for some territory.

The post argues that abstraction is key for embedded agency problems because the underlying challenge of embedded world models is that the agent (the map) is smaller than the environment it is modeling (the territory), and so inherently has to throw some information away.

Some simple questions around abstraction that we might want to answer include:

- Given a map-making process, characterize the queries whose answers the map can reliably predict.

- Given some representation of the map-territory correspondence, translate queries from the territory-representation to the map-representation and vice versa.

- Given a territory, characterize classes of queries which can be reliably answered using a map much smaller than the territory itself.

- Given a territory and a class of queries, construct a map which throws out as much information as possible while still allowing accurate prediction over the query class.

The post argues that once we create the simple theory, we will have a natural way of looking at more challenging problems with embedded agency, like the problem of self-referential maps, the problem of other map-makers, and the problem of self-reasoning

that arises when the produced map includes an abstraction of the map-making process itself.

Asya's opinion: My impression is that embedded agency problems as a class of problems are very young, extremely entangled, and characterized by a lot of confusion. I am enthusiastic about attempts to decrease confusion and intuitively, abstraction does feel like a key component to doing that.

That being said, my guess is that it's difficult to predictably suggest the most promising research directions in a space that's so entangled. For example, one thread in the comments of this post discusses the fact that this theory of abstraction as presented looks at "one-shot" agency where the system takes in some data once and then outputs it, rather than "dynamic" agency where a system takes in data and outputs decisions repeatedly over time. Abram Demski argues that the "dynamic" nature of embedded agency is a central part of the problem and that it may be more valuable and neglected to put research emphasis there.

Dissolving Confusion around Functional Decision Theory *(Stephen Casper)* (summarized by Rohin): This post argues for functional decision theory (FDT) on the basis of the following two principles:

1. Questions in decision theory are not about what "choice" you should make with your "free will", but about what source code you should be running.

2. P "subjunctively depends" on A to the extent that P's predictions of A depend on correlations that can't be confounded by choosing the source code that A runs.

Rohin's opinion: I liked these principles, especially the notion that subjunctive dependence should be cashed out as "correlations that aren't destroyed by changing the source code". This isn't a perfect criterion: FDT can and should apply to humans as well, but we *don't* have control over our source code.

Predictors exist: CDT going bonkers... forever *(Stuart Armstrong)* (summarized by Rohin): Consider a setting in which an agent can play a game against a predictor. The agent can choose to say zero or one. It gets 3 utility if it says something different from the predictor, and -1 utility if it says the same thing. If the predictor is near-perfect, but the agent models its actions as independent of the predictor (since the prediction was made in the past), then the agent will have some belief about the prediction and will choose the less likely action for expected utility at least 1, and will continually lose.

ACDT: a hack-y acausal decision theory *(Stuart Armstrong)* (summarized by Rohin): The problem with the previous agent is that it never learns that it has the wrong causal model. If the agent is able to learn a better causal model from experience, then it can learn that the predictor can actually predict the agent successfully, and so will no longer expect a 50% chance of winning, and it will stop playing the game.

Miscellaneous (Alignment)

Clarifying The Malignity of the Universal Prior: The Lexical Update *(interstice)*

Other progress in AI

Reinforcement learning

Reward-Conditioned Policies *(Aviral Kumar et al)* (summarized by Nicholas): Standard RL algorithms create a policy that maximizes a reward function; the *Reward-Conditioned Policy* algorithm instead creates a policy that can achieve a particular reward value passed in as an input. This allows the policy to be trained via supervised regression on a dataset. Each example in the dataset consists of a state, action, and either a return or an advantage, referred to as *Z*. The network then predicts the action based on the state and *Z*. The learned model is able to generalize to policies for larger returns. During training, the target value is sampled from a distribution that gradually increases so that it continues to learn higher rewards.

During evaluation, they then feed in the state and a high target value of *Z* (set one standard deviation above the average in their paper.) This enables them to achieve solid - but not state of the art - performance on a variety of the OpenAI Gym benchmark tasks. They also run ablation studies showing, among other things, that the policy is indeed accurate in achieving the target reward it aims for.

Nicholas's opinion: One of the dangers of training powerful AI to maximize a reward function is that optimizing the function to extreme values may no longer correlate with what we want, as in the classic paperclip maximizer example. I think RCP provides an interesting solution to that problem; if we can instead specify a good, but reasonable, value, we may be able to avoid those extreme cases. We can then gradually increase the desired reward without retraining while continuously monitoring for issues. I think there are likely flaws in the above scheme, but I am optimistic in general about the potential of finding alternate ways to communicate goals to an agent.

One piece I am still curious about is whether the policy remembers how to achieve lower rewards as its training dataset updates towards higher rewards. They show in a heatmap that the target and actual rewards do match up well, but the target rewards are all sampled quite near each other; it would be interesting to see how well the final policy generalizes to the entire spectrum of target rewards.

Reinforcement Learning Upside Down: Don't Predict Rewards -- Just Map Them to Actions and Training Agents using Upside-Down Reinforcement Learning *(Juergen Schmidhuber)* (summarized by Zach): It's a common understanding that using supervised learning to solve RL problems is challenging because supervised learning works directly with error signals while RL only has access to evaluation signals. The approach in these papers introduce 'upside-down' reinforcement learning (UDRL) as a way to bridge this gap. Instead of learning how to predict rewards, UDRL learns how to take actions when given a state and a desired reward. Then, to get good behavior, we simply ask the policy to take actions that lead to particularly high rewards. The main approach is to slowly increase the desired goal behavior as the agent learns in order to maximize agent performance. The authors evaluate UDRL on the Lunar Lander and the Take Cover environments. UDRL ultimately performs worse on Lunar Lander and better on Take Cover so it's unclear whether or not UDRL is an improvement over popular methods. However, when rewards are made to be sparse UDRL is able to significantly outperform other RL methods.

Zach's opinion: This approach fits neatly with older work including "Learning to Reach Goals" and more recent work such as Hindsight experience replay and Goal-Conditioned Policies. In particular, all of these methods seem to be effective at addressing the difficulty that comes with working with sparse rewards. I also found

myself justifying the utility of selecting the objective of 'learning to achieve general goals' to be related to the idea that seeking power is instrumentally convergent (AN #78).

Rohin's opinion: Both this and the previous paper have explored the idea of conditioning on rewards and predicting actions, trained by supervised learning. While this doesn't hit state-of-the-art performance, it works reasonably well for a new approach.

Planning with Goal-Conditioned Policies *(Soroush Nasiriany, Vitchyr H. Pong et al)* (summarized by Zach): Reinforcement learning can learn complex skills by interacting with the environment. However, temporally extended or long-range decision-making problems require more than just well-honed reactions. **In this paper, the authors investigate whether or not they can obtain the benefits of action planning found in model-based RL without the need to model the environment at the lowest level.** The authors propose a model-free planning framework that learns low-level goal-conditioned policies that use their value functions as implicit models. Goal-conditioned policies are policies that can be trained to reach a goal state provided as an additional input. Given a goal-conditioned policy, the agent can then plan over intermediate subgoals (goal states) using a goal-conditioned value function to estimate reachability. Since the state space is large, the authors propose what they call latent embeddings for abstracted planning (LEAP), which is able to find useful subgoals by first searching a much smaller latent representation space and then planning a sequence of reachable subgoals that reaches the target state. In experiments, LEAP significantly outperforms prior algorithms on 2D navigation and push/reach tasks. Moreover, their method can get a quadruped ant to navigate around walls which is difficult because much of the planning happens in configuration space. This shows that LEAP is able to be extended to non-visual domains.

Zach's opinion: The presentation of the paper is clear. In particular, the idea of planning a sequence of maximally feasible subgoals seems particularly intuitive. In general, I think that LEAP relies on the clever idea of reusing trajectory data to augment the data-set for the goal-conditioned policy. As the authors noted, the question of exploration was mostly neglected. I wonder how well the idea of reusing trajectory data generalizes to the general exploration problem.

Rohin's opinion: The general goal of inferring hierarchy and using this to plan more efficiently seems very compelling but hard to do well; this is the goal in most hierarchical RL algorithms and Learning Latent Plans from Play (AN #65).

Dream to Control: Learning Behaviors by Latent Imagination *(Danijar Hafner et al)* (summarized by Cody): In the past year or so, the idea of learning a transition model in a latent space has gained traction, motivated by the hope that such an approach could combine the best of the worlds of model-free and model-based learning. The central appeal of learning a latent transition model is that it allows you to imagine future trajectories in a potentially high-dimensional, structured observation space without actually having to generate those high-dimensional observations.

Dreamer builds on a prior model by the same authors, PlaNet (AN #33), which learned a latent representation of the observations, p(s|o), trained both through a VAE-style observation reconstruction loss, and also a transition model q(s-next|s, a), which is trained to predict the state at the next step given only the state at the prior one, with no next-step observation data. Together, these two models allow you to simulate action-conditioned trajectories through latent state space. If you then predict reward

from state, you can use this to simulate the value of trajectories. Dreamer extends on this by also training an Actor Critic-style model on top of states to predict action and value, forcing the state representation to not only capture next-step transition information, but also information relevant to predicting future rewards. The authors claim this extension makes their model more able to solve long-horizon problems, because the predicted value function can capture far-future rewards without needing to simulate the entire way there. Empirically, there seems to be reasonable evidence that this claim plays out, at least within the fairly simple environments the model is tested in.

Cody's opinion: The extension from PlaNet (adding actor-critic rather than direct single-step reward prediction) is relatively straightforward, but I think latent models are an interesting area - especially if they eventually become at all possible to interpret - and so I'm happy to see more work in this area.

How to Identify an Immoral Maze

Previously in sequence: Moloch Hasn't Won, Perfect Competition, Imperfect Competition, Does Big Business Hate Your Family?, What is Life in an Immoral Maze?, Stripping Away the Protections, What is Success in an Immoral Maze?

Immoral mazes (hereafter mazes), as laid out in the book *Moral Mazes*, are toxic organizations. Working for them puts tremendous pressure on you to prioritize getting ahead in the organization over everything else. Middle managers are particularly affected – they are pushed to sacrifice not only all of their time, but also things such as their morality, family and ability to think clearly. Only those who go all-in doing this get ahead, and even most of them fail.

Even successfully getting ahead is little consolation.

Mazes exert similar pressures on those who do business with them or work in non-managerial roles, to a lesser but substantial degree.

The best defense is to identify mazes before you agree to work for or do business with them, and choose to work or do business elsewhere. At a minimum, one's eyes should be open, and the costs involved must be fully factored in before making such decisions.

This makes it important to figure out what parts of what organizations are mazes, and to what extent. This is hard to get exactly right.

What is easier is using simple heuristics to get a good approximation, then keeping an eye out for and updating on new evidence.

I offer seven heuristics, the first two of which will do the bulk of the work on their own. You benefit from the 'right' answer to all of them even absent concerns about mazes, so they are good questions to get into the habit of asking.

1. How many levels of hierarchy exist?

Full mazes require *at least* three levels of hierarchy, without which one cannot have middle management.

Each level beyond that makes things worse. The fourth and fifth levels both make things much worse.

With only one level, there's nothing to worry about.

With only two levels, a boss and those who report to the boss, the boss has skin in the game, no boss causing problems for them, and not enough reason to reward bad outcomes.

With three levels, there are middle managers in the second layer, so one should be wary. But things are unlikely to be too bad. No middle manager has a boss or underling who is *also* a middle manager. This means that in any interaction between non-equals either involves the head of the company, or it involves someone 'on the line' who doesn't have anyone reporting to them, and must deal with object-level

reality. Either of them has reason to keep things grounded. Since there is only one person at the top, every conversation includes someone who interacts regularly with object level reality.

With four levels, we start to have interactions between middle managers in charge of each other. These dynamics start to get serious, but everyone still interacts with someone on the top or bottom.

At five levels, we have people who never interact directly with *either* the boss *or* anyone dealing with the object level.

At six levels, those people interact with each other.

And so on.

Meanwhile, the boss has less and less need or ability to comprehend the object level, and we get more and more problems with lack of skin in the game, which is question two.

At least one of the corporations in *Moral Mazes* had more than *twenty* ranks. That is *way, way* too many. By that point, it would be surprising if you weren't doomed. I have *actual no idea* how to have twenty ranks and keep things sane.

Note that those outside the company, such as investors or regulators, seem like they should effectively count as a level under some circumstances, but not under others.

As a spot check, I looked back on the jobs I've had. This matches my experience.

Most impressive is that I can observe what happened when several of those jobs *added new layers of hierarchy.* This led *in every case* to traceable ways to additional maze-like behavior. In every case, that made life much worse for me and other employees, and hurt our productivity. In one case I was running the company at the time, and it still happened.

I would be very wary of any organization that had four levels of hierarchy. I would be progressively more skeptical of any organization with more than that, to the point of assuming it was a maze until proven otherwise.

2. Do people have skin in the game?

Skin in the game is a robust defense against mazes, if it can be distributed widely enough and in the right ways. That can be tough. There's only 100% total equity to go around.

One can only reward what can be observed or often only what can be quantified and measured. Something about Goodhart's Law, and so on. The problem with levels of hierarchy and middle management is in large part a problem of inability to provide skin in the game.

For sufficiently large organizations, as described in *Moral Mazes,* skin in the game is not so much spread thin as deliberately destroyed. The successful keep enough momentum to run away from the consequences of their problems. This alone is fatal.

If an organization has solved these problems for real, it likely isn't a maze.

If an organization lacks skin in the game and also has many levels of hierarchy, you're almost certainly dealing with a maze.

If it lacks skin in the game but also lacks levels of hierarchy, maze levels can differ. But also keep in mind that lack of skin in the game causes a whole host of problems. Only some of those are the problems of mazes. Detailing these issues is beyond the scope here, but be highly skeptical whenever skin in the game is lacking.

3. Do people have soul in the game?

What's better than having skin in the game? Having soul in the game. Caring deeply about the outcome for reasons other than money, or your own liability, or being potentially scapegoated. Caring for existential reasons, not commercial ones.

Soul in the game is incompatible with mazes. Mazes will eliminate anyone with soul in the game. Therefore, if the people you work for have soul in the game, you're safe. If you have it too, you'll be a lot happier, and likely doing something worthwhile. Things will be much better on most fronts.

It's worth prioritizing soul in the game, above and beyond skin in the game.

4. How do people describe their job when you ask?

Remember this quote:

> *When managers describe their work to an outsider, they almost always first say: "I work for [Bill James]" or "I report to [Harry Mills]" or "I'm in [Joe Bell's] group,"* and only then proceed to describe their actual work functions. (Location 387, Quote 2)*

You want them to say *almost anything else*. Anything that does not make you recoil in horror a different way. Hopefully something worthwhile and interesting. I don't know how good this rule is, but I suspect it's quite powerful.

5. Is there diversity of skill levels? Is excellence possible and rewarded?

The belief that all middle managers have the same skills, and are all equally capable of doing any managerial job aside from the politics involved, is a lot of what makes mazes so bad. If there is no *good* reason to diverge from standard practice, if everybody knows that you *cannot do better,* then any divergence is blameworthy, and shows you are not doing your job. There's no need to ask why, or what advantages it might have.

It also all but ensures the wrong answer to the next question.

6. Is there slack?

A world without slack is not a place one wants to be. Mazes systematically erase all slack. Slack is evidence of not being fully committed, and given that everyone's skills

are equal and competition is perfect, holding anything back means losing even if undetected.

7. Pay Attention

Sounds silly, but it works. Observe people and what they do and how they do it. If you work in a maze for long enough, you're not going to shout it from the rooftops, but every sentence you speak will reflect it.

And as always, when people tell you who they are, believe them.

Other Notes

These questions do not differentiate between corporations, non-profits, governments, parties, clubs or other organizational forms. That's not a good indicator. Corporations are only the original observed case.

Asking how proposed or expected changes will change the answers to these questions is a good way to know if those changes will raise the maze level of an organization.

Like most puzzles, there are multiple solutions, and the pieces reinforce each other. Most of the time, hardcore mazes will give alarm-bell level answers to all seven heuristics.

Are there any other good simple heuristics?

Next is How to Work With Moral Mazes, providing my best advice in detail to those dealing with the threat of mazes on a personal level.

Moral uncertainty: What kind of 'should' is involved?

This post follows on from my prior post; consider reading that post first.

We are often forced to make decisions under conditions of uncertainty. This may be empirical uncertainty (e.g., what is the likelihood that nuclear war would cause human extinction?), or it may be moral uncertainty (e.g., does the wellbeing of future generations matter morally?).

In my prior post, I discussed overlaps with and distinctions between moral uncertainty and related concepts. In this post, I continue my attempt to clarify **what moral uncertainty actually *is*** (rather than how to make decisions when morally uncertain, which is covered later in the sequence). Specifically, here I'll discuss:

1. Is what we "ought to do" (or "should do") under moral uncertainty an *objective* or *subjective* (i.e., *belief-relative*) matter?
2. Is what we "ought to do" (or "should do") under moral uncertainty a matter of *rationality* or *morality*?

An important aim will be simply clarifying the questions and terms themselves. That said, to foreshadow, the tentative "answers" I'll arrive at are:

1. It seems both more intuitive and more action-guiding to say that the "ought" is *subjective*.
2. Whether the "ought" is a rational or a moral one may be a "merely verbal" dispute with no practical significance. But I'm very confident that interpreting the "ought" as a matter of *rationality* works in any case (i.e., whether or not interpreting it as a matter of *morality* does, and whether or not the distinction really matters).

This post *doesn't* explicitly address what types of moral uncertainty would be *meaningful for moral antirealists and/or subjectivists*; I discuss that topic in a separate post.[1]

*Epistemic status: The concepts covered here are broad, fuzzy, and overlap in various ways, making definitions and distinctions between them almost inevitably debatable. Additionally, I'm not an expert in these topics (though I have now spent a couple weeks mostly reading about them). I've tried to mostly collect, summarise, and synthesise **existing** ideas (from academic philosophy and the LessWrong and EA communities). I'd appreciate feedback or comments in relation to any mistakes, unclear phrasings, etc. (and just in general!).*

Objective or subjective?

*(Note: What I discuss here is **not** the same as the objectivism vs subjectivism debate **in metaethics**.)*

As I noted in a prior post:

> *Subjective* normativity relates to what one should do *based on what one believes*, whereas *objective* normativity relates to what one "actually" should do (i.e., based on the true state of affairs).

Hilary Greaves & Owen Cotton-Barratt give an example of this distinction in the context of *empirical* uncertainty:

> Suppose Alice packs the waterproofs but, as the day turns out, it does not rain. Does it follow that Alice made the wrong decision? In one (**objective**) sense of "wrong", yes: thanks to that decision, she experienced the mild but unnecessary inconvenience of carrying bulky raingear around all day. But in a second (more **subjective**) sense, clearly it need not follow that the decision was wrong: if the probability of rain was sufficiently high and Alice sufficiently dislikes getting wet, her decision could easily be the appropriate one to make given her state of ignorance about how the weather would in fact turn out. Normative theories of decision-making under uncertainty aim to capture this second, more subjective, type of evaluation; the standard such account is expected utility theory.

Greaves & Cotton-Barratt then make the analogous distinction for *moral* uncertainty:

> How should one choose, when facing relevant **moral** uncertainty? In one (**objective**) sense, of course, what one should do is simply what the true moral hypothesis says one should do. But it seems there is also a second sense of "should", analogous to the **subjective** "should" for empirical uncertainty, capturing the sense in which it is appropriate for the agent facing moral uncertainty to be guided by her moral credences [i.e., beliefs], whatever the moral facts may be. (emphasis added)

(This objective vs subjective distinction seems to me somewhat similar - though not identical - to the distinction between *ex post* and *ex ante* thinking. We might say that Alice made the right decision *ex ante* - i.e., based on what she knew when she made her decision - even if it *turned out* - *ex post* - that the other decision would've worked out better.)

MacAskill notes that, in both the empirical and moral contexts, "The principal argument for thinking that there must be a subjective sense of 'ought' is because the objective sense of 'ought' is not sufficiently action-guiding." He illustrates this in the case of moral uncertainty with the following example:

> Susan is a doctor, who faces three sick individuals, Greg, Harold and Harry. Greg is a human patient, whereas Harold and Harry are chimpanzees. They all suffer from the same condition. She has a vial of a drug, D. If she administers all of drug D to Greg, he will be completely cured, and if she administers all of drug to the chimpanzees, they will both be completely cured (health 100%). If she splits the drug between the three, then Greg will be almost completely cured (health 99%), and Harold and Harry will be partially cured (health 50%). She is unsure about the value of the welfare of non-human animals: she thinks it is equally likely that chimpanzees' welfare has no moral value and that chimpanzees' welfare has the same moral value as human welfare. And, let us suppose, there is no way that she can improve her epistemic state with respect to the relative value of humans and chimpanzees.
>
> [...]
>
> Her three options are as follows:

A: Give all of the drug to Greg

B: Split the drug

C: Give all of the drug to Harold and Harry

Her decision can be represented in the following table, using numbers to represent how good each outcome would be.

	Chimpanzee welfare is of no moral value – 50%	Chimpanzee welfare is of significant moral value – 50%
A	100	0
B	99	199
C	0	200

Finally, suppose that, according to the true moral theory, chimpanzee welfare is of the same moral value as human welfare and that therefore, she should give all of the drug to Harold and Harry. What should she do?

Clearly, the best *outcome* would occur if Susan does C. But she doesn't *know* that that would cause the best outcome, because she doesn't *know* what the "true moral theory" is. She thus has no way to act on the advice "Just do what is *objectively* morally right." Meanwhile, as MacAskill notes, "it seems it would be morally reckless for Susan **not** to choose option B: given what she knows, she would be risking severe wrongdoing by choosing either option A or option C" (emphasis added).

To capture the intuition the Susan should choose option B, and **to provide *actually followable* guidance for action, we need to accept that there is a *subjective* sense of "should"** (or of "ought") - a sense of "should" that *depends in part on what one believes*. (This could also be called a "belief-relative" or "credence-relative" sense of "should".)[2]

An additional argument in favour of accepting that there's a subjective "should" in relation to moral uncertainty is consistency with how we treat *empirical* uncertainty, where most people accept that there's a subjective "should".[3] This argument is made regularly, including by MacAskill and by Greaves & Cotton-Barratt, and it seems particularly compelling when one considers that it's often difficult to draw clear lines between empirical and moral uncertainty (see my prior post). That is, if it's often hard to say whether an uncertainty is empirical or moral, it seems strange to say we should accept a subjective "should" under empirical uncertainty but *not* under moral uncertainty.

Ultimately, most of what I've read on moral uncertainty is premised on there being a subjective sense of "should", and much of this sequence will rest on that premise also.[4] As far as I can tell, this seems necessary if we are to come up with any meaningful, action-guiding approaches for decision-making under moral uncertainty ("metanormative theories").

But I should note that *some* writers *do* appear to argue that there's only an objective sense of "should" (one example, I think, is Weatherson, though he uses different language and I've only skimmed his paper). Furthermore, while I can't see how this could lead to action-guiding principles for *making decisions under uncertainty*, it does seem to me that it'd still allow for *resolving* one's uncertainty. In other words, if we do recognise only objective "oughts":

- We may be stuck with fairly useless principles for decision-making, such as "Just do what's actually right, even when you don't know what's actually right"
- But (as far as I can tell) we could still be guided to *clarify* and *reduce* our uncertainties, and thereby bring our beliefs more in line with what's actually right.

Rational or moral?

> There is also debate about what precisely kind of "should" is involved [in cases of moral uncertainty]: rational, moral, or something else again. (Greaves & Cotton-Barratt)

For example, in the above example of Susan the doctor, are we wondering what she *rationally* ought to do, given her moral uncertainty about the moral status of chimpanzees, or what she *morally* ought to do?

It may not matter either way

Unfortunately, even after having read up on this, it's not actually clear to me what the distinction is meant to be. In particular, I haven't come across a clear explanation of what it would mean for the "should" or "ought" to be *moral*. I suspect that what that would mean would be partly a matter of interpretation, and that some definitions of a "moral" should could be effectively the same as those for a "rational" should. (But I should note that I didn't look exhaustively for such explanations and definitions.)

Additionally, both Greaves & Cotton-Barratt and MacAskill explicitly avoid the question of whether what one "ought to do" under moral uncertainty is a matter of rationality or morality.[5] This does not seem to at all hold them back from making valuable contributions to the literature on moral uncertainty (and, more specifically, on how to make decisions when morally uncertain).

Together, the above points make me **inclined to believe (though with low confidence) that this may be a "merely verbal" debate with no real, practical implications** (at least while the words involved remain as fuzzy as they are).

However, I still did come to two less-dismissive conclusions:

1. I'm very confident that the project of working out meaningful, action-guiding principles for decision-making under moral uncertainty **makes sense if we see the relevant "should" as a rational one**. (Note: This doesn't mean that I think the "should" *has* to be seen as a rational one.)
2. I'm less sure whether that project would make sense if we see the relevant "should" as a moral one. (Note: This doesn't mean I have any particular reason to believe it *wouldn't* make sense if we see the "should" as a moral one.)

I provide my reasoning behind these conclusions below, though, given my sense that this debate may lack practical significance, **some readers may wish to just skip to the next section**.

A rational "should" likely works

Bykvist writes:

> An alternative way to understand the ought relevant to moral uncertainty is in terms of rationality (MacAskillet al., forthcoming; Sepielli, 2013). Rationality, in one important sense at least, has to do with what one should do or intend, given one's beliefs and preferences. This is the kind of rationality that decision theory often is seen as invoking. It can be spelled out in different ways. One is to see it as a matter of coherence: It is rational to do or intend what coheres with one's beliefs and preferences (Broome, 2013; for a critic, see Arpaly, 2000). Another way to spell it out is to understand it as matter of rational processes: it is rational to do or intend what would be the output of a rational process, which starts with one's beliefs and preferences (Kolodny, 2007).
>
> To apply the general idea to moral uncertainty, we do not need to take stand on which version is correct. We only need to assume that when a conscientious moral agent faces moral uncertainty, she cares about doing right and avoid doing wrong but is uncertain about the moral status of her actions. She prefers doing right to doing wrong and is indifferent between different right doings (at least when the right doings have the same moral value, that is, none is morally supererogatory). She also cares more about serious wrongdoings than minor wrongdoings. The idea is then to apply traditional decision theoretical principles, according to which rational choice is some function of the agent's preferences (utilities) and beliefs (credences). Of course, different decision-theories provide different principles (and require different kinds of utility information). But the plausible ones at least agree on cases where one option dominates another.
>
> Suppose that you are considering only two theories (which is to simplify considerably, but we only need a logically possible case): "business as usual," according to which it is permissible to eat factory-farmed meat and permissible to eat vegetables, and "vegetarianism," according to which it is impermissible to eat factory-farmed meat and permissible to eat vegetables. Suppose further that you have slightly more confidence in "business as usual." The option of eating vegetables will dominate the option of eating meat in terms of your own preferences: No matter which moral theory is true, by eating vegetables, you will ensure an outcome that you weakly [prefer] to the alternative outcome: if "vegetarianism" is true, you prefer the outcome; if "business as usual is true," you are indifferent between the outcomes. The rational thing for you to do is thus to eat vegetables, given your beliefs and preferences. (lines breaks added)

It seems to me that that reasoning makes perfect sense, and that we can have valid, meaningful, action-guiding principles about what one *rationally* (and subjectively) should do given one's moral uncertainty. This seems further supported by the approach Christian Tarsney takes, which seems to be useful and to also treat the relevant "should" as a rational one.

Furthermore, MacAskill seems to suggest that there's a correlation between (a) writers fully engaging with the project of working out action-guiding principles for decision-

making under moral uncertainty and (b) writers considering the relevant "should" to be rational (rather than moral):

> (Lockhart 2000, 24,26), (Sepielli 2009, 10) and (Ross 2006) all take metanormative norms to be norms of rationality. (Weatherson 2014) and (Harman 2014) both understand metanormative norms as moral norms. So there is an odd situation in the literature where the defenders of metanormavism (Lockhart, Ross, and Sepielli) and the critics of the view (Weatherson and Harman) seem to be talking past one another.

A moral "should" may or may not work

I haven't seen any writer (a) *explicitly* state that they understand the relevant "should" to be a moral one, and then (b) go on to fully engage with the project of working out meaningful, action-guiding principles for decision-making under moral uncertainty. Thus, I have an absence of evidence that one can engage in that project while seeing the "should" as moral, and I take this as (very weak) evidence that one can't engage in that project while seeing the "should" that way.

Additionally, as noted above, MacAskill writes that Weatherson and Harman (who seem fairly dismissive of that project) see the relevant "should" as a moral one. Arguably, this is evidence that that project of finding such action-guiding principles won't make sense if we see the "should" as moral (rather than rational). However, I consider this to *also* be very weak evidence, because:

- It's only two data points.
- It's just a correlation anyway.
- I haven't closely investigated the "correlation" myself. That is, I haven't checked whether or not Weatherson and Harman's reasons for dismissiveness seem highly related to them seeing the "should" as moral rather than rational.

Closing remarks

In this post, I've aimed to:

- Clarify what is meant by the question "Is what we "ought to do" under moral uncertainty is an *objective* or *subjective* matter?"
- Clarify what is meant by the question "Is that 'ought' a matter of *rationality* or of *morality*?"
- Argue that it seems both more intuitive and more action-guiding to say that the "ought" is *subjective*.
- Argue that whether the "ought" is a rational or a moral one may be a "merely verbal" dispute with no practical significance (but that interpreting the "ought" as a matter of *rationality* works in any case).

I hope this has helped give readers more clarity on the seemingly neglected matter of what we actually *mean* by moral uncertainty. (And as always, I'd welcome any feedback or comments!)

My next posts will continue in a similar vein, but this time building to the question of whether, when we're talking about moral uncertainty, we're actually talking about *moral risk* rather than about *moral (Knightian) uncertainty* - and whether such a

distinction is truly meaningful. (To do so, I'll first discuss the risk-uncertainty distinction *in general*, and the related matter of unknown unknowns, before applying these ideas in the context of *moral* risk/uncertainty in particular.)

---

1. But the current post is still relevant for many types of moral antirealist. As noted in my last post, this sequence will sometimes use language that may appear to endorse or presume moral realism, but this is essentially just for convenience. ↩

2. We could further divide subjective normativity up into, roughly, "what one should do based on what one *actually* believes" and "what one should do based on what it *would be reasonable* for one to believe". The following quote, while not directly addressing that *exact* distinction, seems relevant:

   > Before moving on, we should distinguish subjective credences, that is, degrees of belief, from epistemic credences, that is, the degree of belief that one is epistemically justified in having, given one's evidence. When I use the term 'credence' I refer to epistemic credences (though much of my discussion could be applied to a parallel discussion involving subjective credences); when I want to refer to subjective credences I use the term 'degrees of belief'.
   >
   > The reason for this is that appropriateness seems to have some sort of normative force: if it is most appropriate for someone to do something, it seems that, other things being equal, they ought, in the relevant sense of 'ought', to do it. But people can have crazy beliefs: a psychopath might think that a killing spree is the most moral thing to do. But there's no sense in which the psychopath ought to go on a killing spree: rather, he ought to revise his beliefs. We can only capture that idea if we talk about epistemic credences, rather than degrees of belief.

   (I found that quote in this comment, where it's attributed to MacAskill's *BPhil* thesis. Unfortunately, I can't seem to access that thesis, including via Wayback Machine.) ↩

3. Though note that Greaves and Cotton-Barratt write:

   > Not everyone does recognise a subjective reading of the moral 'ought', even in the case of empirical uncertainty. One can distinguish between objectivist, (rational-)credence-relative and pluralist views on this matter. According to objectivists (Moore, 1903; Moore, 1912; Ross, 1930, p.32; Thomson, 1986, esp. pp. 177-9; Graham, 2010; Bykvist and Olson, 2011) (respectively, credence-relativists (Prichard, 1933; Ross, 1939; Howard-Snyder, 2005; Zimmermann, 2006; Zimmerman, 2009; Mason, 2013), the "ought" of morality is uniquely an objective (respectively, a credence-relative) one. According to pluralists, "ought" is ambiguous between these two readings (Russell, 1966; Gibbard, 2005; Parfit, 2011; Portmore, 2011; Dorsey, 2012; Olsen, 2017), or varies between the two readings according to context (Kolodny and Macfarlane, 2010).

   ↩

4. In the following quote, Bykvist provides what seems to me (if I'm interpreting it correctly) to be a different way of explaining something similar to the objective vs subjective distinction.

> One possible explanation of why so few philosophers have engaged with moral uncertainty might be serious doubt about whether it makes much sense to ask about what one ought do when one is uncertain about what one ought to do. The obvious answer to this question might be thought to be: "you ought to do what you ought to do, no matter whether or not you are certain about it" (Weatherson, 2002, 2014). However, this assumes the same sense of "ought" throughout.
>
> A better option is to assume that there are different kinds of moral ought. We are asking what we morally ought to do, in one sense of ought, when we are not certain about what we morally ought to do, in another sense of ought. One way to make this idea more precise is to think about the different senses as different levels of moral ought. When we face a moral problem, we are asking what we morally ought to do, at the first level. Standard moral theories, such as utilitarianism, Kantianism, and virtue ethics, provide answers to this question. In a case of moral uncertainty, we are moving up one level and asking about what we ought to do, at the second level, when we are not sure what we ought to do at the first level. At this second level, we take into account our credence in various hypotheses about what we ought to do at the first level and what these hypotheses say about the moral value of each action (MacAskill et al., forthcoming). This second level ought provides a way to cope with the moral uncertainty at the first level. It gives us a verdict of how to best manage the risk of doing first order moral wrongs. That there is such a second-level moral ought of coping with first-order moral risks seems to be supported by the fact that agents are morally criticizable when they, knowing all the relevant empirical facts, do what they think is very likely to be a first-order moral wrong when there is another option that is known not to pose any risk of such wrongdoing.

Yet another (and I think similar) way of framing this sort of distinction could make use of the following two terms: "A criterion of rightness tells us what it takes for an action to be right (if it's actions we're looking at). A decision procedure is something that we use when we're thinking about what to do" (Askell).

Specifically, we might say that the true first-order moral theory provides objective "criteria of rightness", but that we don't have direct access to what these are. As such, we can use a second-order "decision procedure" that attempts to lead us to take actions that are close as possible to the best actions (according to the unknown criteria of rightness). To do so, this decision procedure must make use of our credences (beliefs) in various moral theories, and is thus subjective. ↩

5. Greaves & Cotton-Barratt write: "For the purpose of this article, we will [...] not take a stand on what kind of "should" [is involved in cases of moral uncertainty]. Our question is how the "should" in question behaves in purely extensional terms. Say that an answer to that question is a *metanormative theory*."

   MacAskill writes: "I introduce the technical term 'appropriateness' in order to remain neutral on the issue of whether metanormative norms are rational norms, or some other sort of norms (though noting that they can't be first-order norms provided by first-order normative theories, on pain of inconsistency)." ↩

Healing vs. exercise analogies for emotional work

I know a fair number of people who put in a lot of effort into things like emotional healing, digging up and dealing with buried trauma, meditative and therapy practices, and so on. (I count myself in this category.)

And I think that there's a thing that sometimes happens when other people see all of this, which is that it all seems kinda fake. I say this because even I have this thought sometimes. The core of the thought is something like, "if all of this stuff really worked, shouldn't you be *finished* sometime? You claim that practice X was really beneficial, so why have you now been talking about the way that practice Y is great – is any of them really that good if you keep jumping between them?"

And there is something to this suspicion. I do think that jumping from thing to thing, each time claiming that you have found something amazing and transformative while you are actually only deluding yourself, *is* definitely a thing that sometimes happens. I can say this because I've been that person, too.

But it's not the only possibility. Sometimes the moving from thing to thing *does* mean that you are getting genuine value out of each, and you work on each until you hit diminishing returns, and then you move on to the next practice to help deal with the issues that the previous one didn't address.

And it's worth noting that to the skeptical mind, the *opposite* pattern can be suspicious too. Sometimes someone does stick with just one practice – a particular style of meditation, say – for years, maybe decades. And keeps talking about how great and healing it is. And again the person who keeps hearing this starts wondering, okay, if it's so healing, why are you not totally healed yet?

And again, there is something to that suspicion. Sometimes people *do* stick to one thing and think that it is amazing, even if it is not really delivering them any results, and they would be better off switching to something else.

But then sometimes it really _is_ the case that their practice just *is* that good, and they keep getting consistent results.

I think that the major issue here is that "healing" isn't quite the right metaphor. Yes, much of what these practices do could be considered healing, in that they can help you resolve old stuff, possibly for good.

But the way we usually conceive of healing is that you have some specific sickness or injury, then it's healed, and then you are healthy and don't need to do any more healing until you get sick again. And that's not quite the right model for these kinds of practices.

I think that a better model would be physical exercise. Just like the emotional practices, exercise can be useful for healing – I am counting physiotherapy as a form of physical exercise here, though obviously exercise can help heal even if it is not explicitly physiotherapy. But even though healing is one of the things that exercise does, that's not its *only* purpose.

If someone said that they had maintained a jogging habit every day for the last twenty years and that it made them feel consistently amazing, nobody would find that particularly suspicious.

And if someone said that they had done yoga for flexibility a while, then taken up running for the cardio, injured themselves and done physiotherapy for a while, and then started doing weightlifting for the sake of muscle, and each of those had been exactly the right thing to do, then that wouldn't be very suspicious either.

A simple "healthy/unhealthy" model isn't any better for mental and emotional well-being than it is for physical shape. There are things that count as genuine injuries and diseases, yes, but there are also things which require active maintenance, as well as different subareas that you may want to focus on. You might stick with the same practices for a long time, your whole life even, if they seem particularly effective. And you may also want to switch practices from time to time, because you no longer need an old one, or in response to new needs from changed circumstances, or just for the sake of variety.

Using Vickrey auctions as a price discovery mechanism

This is a linkpost for https://kevinlynagh.com/notes/pricing-niche-products/

I recommend "Pricing niche products: Why sell a mechanical keyboard kit for $1,668?" for providing a practical case study in price dynamics that helped with my economic intuitions.

The author's friend had created a new custom keyboard kit. Their friend's previous kit had sold out in minutes, so clearly something was amiss with their "estimate costs and premiums and then set a price" approach:

> I'm not a fan of this inside-out approach, for several reasons:
>
> - factors like "brand premium" are inherently subjective — the temptation to compare to others limits potential upside and differentiation
> - picking a (new, higher) price may have reputational downsides (because o_f course_ your customers spend all day in mechanical keyboard chat rooms and may gripe about you "selling out the community")
> - you will second guess yourself regardless of the outcome; either you sell out again (goto 0) or you sell too few and then must live with the shame of having $20k worth of unsold keyboard in your garage
>
> The most compelling argument against simply picking a price, though, is that it *limits how much you can learn about your market*.

Instead, they run a Vickrey auction (or "second-price sealed-bid auction") and find that the demand curve supports 3x the list price they would have chosen:

> I can't overstate the benefits of knowing the demand curve.
>
> In my friend's case, the auction let them sell far above their initial price *and* revealed that the market was deep enough to justify a larger production run.

(I discovered this post via The Prepared, a newsletter that I'd strongly recommend.)

Normalization of Deviance

An important, ongoing part of the rationalist project is to build richer mental models for understanding the world. To that end I'd like to briefly share part of my model of the world that seems to be outside the rationalist cannon in an explicit way, but which I think is known well to most, and talk a bit about how I think it is relevant to you, dear reader. Its name is "normalization of deviance".

If you've worked a job, attended school, driven a car, or even just grew up with a guardian, you've most likely experienced normalization of deviance. It happens when your boss tells you to do one thing but all your coworkers do something else and your boss expects you to do the same as them. It happens when the teacher gives you a deadline but lets everyone turn in the assignment late. It happens when you have to speed to keep up with traffic to avoid causing an accident. And it happens when parents lay down rules but routinely allow exceptions such that the rules might as well not even exist.

It took a much less mundane situation for the idea to crystalize and get a name. Diane Vaughan coined the term as part of her research into the causes of the Challenger explosion, where she described normalization of deviance as what happens when people within an organization become so used to deviant behavior that they don't see the deviance, even if that deviance is actively working against an important goal (in the case of Challenger, safety). From her work the idea has spread to considerations in healthcare, aeronautics, security, and, where I learned about it, software engineering. Along the way the idea has generalized from being specifically about organizations, violations of standard operating procedures, and safety to any situation where norms are so regularly violated that they are replaced by the de facto norms of the violations.

I think normalization of deviance shows up all over the place and is likely quietly happening in your life right now just outside where you are bothering to look. Here's some ways I think this might be relevant to you, and I encourage you to mention more in the comments:

- If you are trying to establish a new habit, regular violations of the intended habit may result in a deviant, skewed version of the habit being adopted.
- If you are trying to live up to an ideal (truth telling, vegetarianism, charitable giving, etc.), regularly tolerating violations of that ideal draws you away from it in a sneaky, subtle way that you may still claim to be upholding the ideal when in fact you are not and not even really trying to.
- If you are trying to establish norms in a community, regularly allowing norm violations will result in different norms than those you intended being adopted.

Those mentioned, my purpose in this post is to be informative, but I know that some of you will read this and make the short leap to treating it as advice that you should aim to allow less normalization of deviance, perhaps by being more scrupulous or less forgiving. Maybe, but before you jump to that, I encourage you to remember the adage about reversing all advice. Sometimes normalized "deviance" isn't so much deviance as an illegible norm that is serving an important purpose and "fixing" it will actually break things or otherwise make things worse. And not all deviance is *normalized* deviance: if you don't leave yourself enough slack you'll likely fail from

trying too hard. So I encourage you to know about normalization of deviance, to notice it, and be deliberate about how you choose to respond to it.

Why a New Rationalization Sequence?

This is the first in a five-post mini-sequence about rationalization, which I intend to post one-per-day. And you may ask, why should we have such a sequence?

What is Rationalization and Why is it Bad?

For those of you just tuning in, rationalization is when you take a conclusion you want to reach and try to come up with a argument that concludes it. The argument looks very similar to one in which you started from data, evaluated as well as you could, and reached this conclusion naturally. Almost always similar enough to fool the casual observer, and often similar enough to *fool yourself.*

If you're deliberately rationalizing for an outside audience, that's out-of-scope for this sequence. All the usual ethics and game theory apply.

But if you're involuntarily rationalizing and fooling yourself, then you've failed at epistemics. And your arts have turned against you. Know a lot about scientific failures? Now you can find them in *all* the studies you didn't like!

Didn't Eliezer Already Do This?

Eliezer wrote the against rationalization sequence back in 2007/8. If you haven't read it, you probably should. It does a good job of describing what rationalization is, how it can happen, and how bad it can be. It does *not* provide a lot of tools for you to use in protecting yourself from rationalization. That's what I'll be focusing on here.

And, besides, if we don't revisit a topic this important every decade or so with new developments, then what is this community for?

Is There Hope?

Periodically, I hear someone give up on logical argument completely. "You can find an argument for anything," they say, "Forget logic. Trust [your gut / tradition / me] instead." Which brushes over the question of whether the proposed alternative is any better. There is no royal road to knowledge.

Still, the question needs answering. If rationalization looks just like logic, can we ever escape Cartesian Doubt?

The Psychiatrist Paradox

A common delusion among grandiose schizophrenics in institutions is that they are themselves psychiatrists. Consider a particularly underfunded mental hospital, in which the majority of people who "know" themselves to be psychiatrists are wrong. No

examination of the evidence will convince them otherwise. No matter how overwhelming, some reason to disbelieve will be found.

Given this, should any amount of evidence suffice to convince you that you are such a psychiatrist?

I am not aware of any resolution to this paradox.

The Dreaming Paradox

But the Psychiatrist Paradox is based on an *absolute* fixed belief and *total* rationalization as seen in theoretically ideal schizophrenics. (How closely do real-world schizophrenics approximate this ideal? That question is beyond the scope of this document.) Let's consider people a little more reality-affiliated: the dreaming.

Given that any evidence of awakeness is a thing that can be dreamed, should you ever be more than 90% confident you're awake? (Assuming 16 hours awake and 2 dreaming in a typical 24 hour period.)

(Boring answer: forget confidence, always *act on the assumption* that you're awake because it's erring on the side of safety. We'll come back to this thought.)

(Also boring: most lucid dreaming enthusiasts report they do find evidence of wakefulness or dreaminess which dreams never forge. Assume you haven't found any for yourself.)

Here's my test: I ask my computer to prime factor a large number (around ten digits) and check it by hand. I can dream many things, but I'm not going to dream that my computer doesn't have the `factor` program, nor will I forget how to multiply. And I *can't* dream that it factored correctly, because I can't factor numbers that big.

You can't outsmart an absolute tendency to rationalize, but you can outsmart a *finite* one. Which, I suspect, is what we mostly have.

A Disclaimer Regarding Authorship

Before I start on the meat of the sequence (in the next post) I should make clear that not all these ideas are mine. Unfortunately, I've lost track of which ones are and which aren't, and of who proposed the ones which aren't. And the ones that aren't original to me have still gone through me enough to not be entirely as their original authors portrayed them.

If I tried to untangle this mess and credit properly, I'd never get this written. So onward. If you wish to fix some bit of crediting, leave a comment and I'll try to do something sensible.

Beyond Rationalization

Much of what appears here also applies to ordinary mistakes of logic. I'll try to tag such as they go.

The simplest ideal of thinking deals extensively with uncertainty of external facts, but trusts its own reasoning implicitly. Directly imitating this, when your own reasoning is not 100% trustworthy, is a bad plan. Hopefully this sequence will provide some alternatives.

Next: Red Flags for Rationalization

Update on Ought's experiments on factored evaluation of arguments

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://ought.org/updates/2020-01-11-arguments

Ought has written a detailed update and analysis of recent experiments on factored cognition. These are experiments with human participants and don't involve any machine learning. The goal is to learn about the viability of IDA, Debate, and related approaches to AI alignment. For background, here are some prior LW posts on Ought: Ought: Why it Matters and How to Help, Factored Cognition presentation.

Here is the opening of the research update:

> **Evaluating Arguments One Step at a Time**
>
> We're studying factored cognition: under what conditions can a group of people accomplish complex cognitive tasks if each person only has minimal context?
>
> In a recent experiment, we focused on dividing up the task of evaluating arguments. We created short, structured arguments for claims about movie reviews. We then tried to distinguish valid from invalid arguments by showing each participant only one step of the argument, not the review or the other steps.
>
> In this experiment, we found that:
>
> 1. Factored evaluation of arguments can distinguish some valid from invalid arguments by identifying implausible steps in arguments for false claims.
>
> 2. However, experiment participants disagreed a lot about whether steps were valid or invalid. This method is therefore brittle in its current form, even for arguments which only have 1–5 steps.
>
> 3. More diverse argument and evidence types (besides direct quotes from the text), larger trees, and different participant guidelines should improve results.
>
> In this technical progress update, we describe these findings in depth.

The rest of the post is here.

Safety regulators: A tool for mitigating technological risk

Crossposted to the Effective Altruism Forum

So far the idea of differential technological development has been discussed in a way that either (1) emphasizes ratios of progress rates, (2) ratios of remaining work, (3) maximizing or minimizing correlations (for example, minimizing the overlap between the capability to do harm and the desire to do so), (4) implementing safe tech before developing and implementing unsafe tech, and (5) the occasional niche analysis (possibly see also a complementary aside relating differential outcomes to growth rates in the long run). I haven't seen much work talking about how various capabilities (a generalization of technology) may interact with each other in general in ways that prevent downside effects (though see also The Vulnerable World Hypothesis), and I wish to elaborate on this interaction type.

As technology improves, our *capacity* to do both harm and good increases and each additional capacity unlocks new capacities that can be implemented. For example the invention of engines unlocked railroads, which in turn unlocked more efficient trade networks. However, the invention of engines also enabled the construction of mobile war vehicles. How, in an ideal world, could we implement capacities so we get the outcomes we want while creating minimal harm and risks in the process?

What does implementing a capacity do? It enables us to change something. A normal progression is:

1. We have no control over something (e.g. We cannot generate electricity)
2. We have control but our choices are noisy and partially random (e.g. We can produce electric sparks on occasion but don't know how to use them)
3. Our choices are organized but there are still downside effects (e.g. We can channel electricity to our homes but occasionally people get electrocuted or fires are started)
4. Our use of the technology mostly doesn't have downside effects (e.g. We have capable safety regulators (e.g. insulation, fuses,...) that allows us to minimize fire and electrocution risks)

The problem is that downside effects in stages 2 and 3 could overwhelm the value achieved during those stages and at stage 4, especially when considering powerful game changing technologies that could lead to existential risks.

Even more fundamentally, as agents in the world we want to avoid shifting the expected utility in a negative direction relative to other options (the opportunity costs). We want to implement new capacities in the best sequence, like with any other plan, so as to maximize the value we achieve. The value is a property of an entire plan and the value is harder to think about than just what is the optimal (or safe) next thing to do (ignoring what is done after). We wish to make choosing which capacities to develop more manageable and easier to think about. One way to do this is to make sure that each capacity we implement is immediately an improvement relative to the state we're in before implementing it (this simplification is an example of a greedy algorithm heuristic). What does this simplification imply about the sequence of implementing capacities?

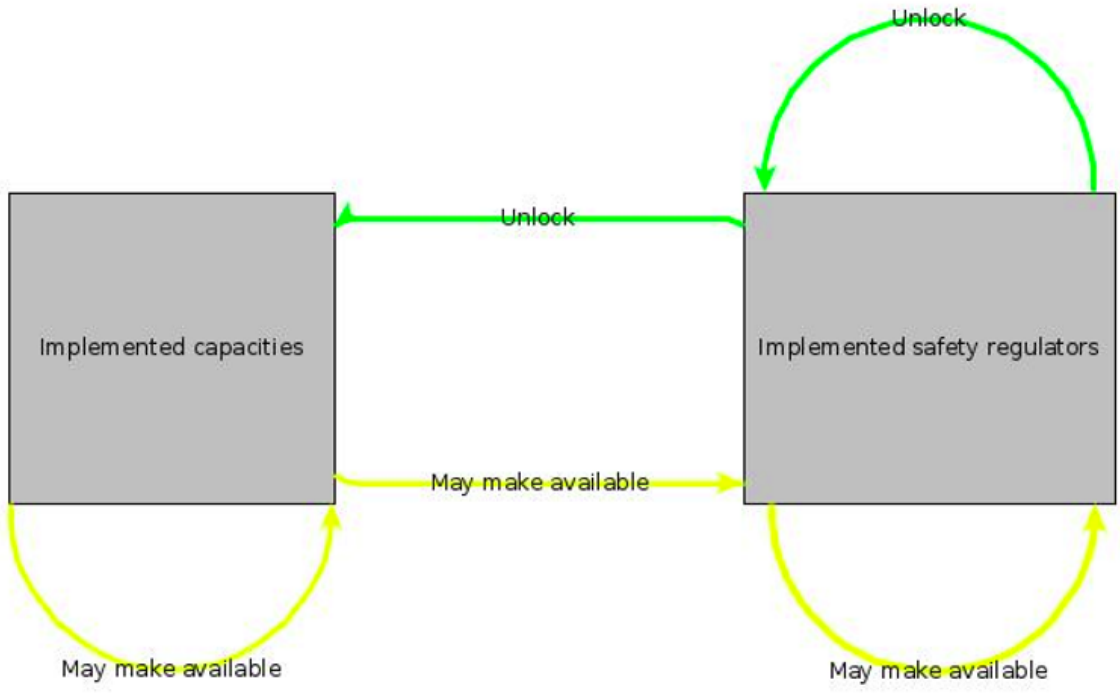
This implies that what we want to do is to have the capacities so we may do good without the downside effects and risks of those capacities. How do we do this? If we're lucky the capacity itself has no downside risks, and we're done. But if we're not lucky we need to implement a regulator on that capacity: a safety regulator. Let's define a *safety regulator* as a capacity that helps control other capacities to mitigate their downside effects. Once a

capacity has been fully safety regulated, it is then unlocked and we can implement it to positive effect.

Some distinctions we want to pay attention to are then:

- A *capacity* - a technology, resource, or plan that changes the world either autonomously or by enabling us to use it
- An *implemented capacity* - a capacity that is implemented
- An *available capacity* - a capacity that can be implemented immediately
- An *unlocked capacity* - a capacity that is safe and beneficial to implement given the technological context, and is also available
- A *potential capacity* - the set of all possible capacities: those already implemented, those being worked on, those that are available and those that exists in theory but need prerequisite capacities to be implemented first.
- A *safety regulator* - a capacity that unlocks other capacities, by mitigating downside effects and possibly providing a prerequisite. (The safety regulator may or may not be unlocked itself at this stage - you may need to implement other safety regulators or capacities to unlock it). Generally, safety regulators are somewhat specialized for the specific capacities they unlock.

Running the suggested heuristic strategy then looks like: If a capacity is unlocked, then implement it; otherwise, implement either an unlocked safety regulator for it first or choose a different capacity to implement. We could call this a *safety regulated capacity expanding feedback loop*. For instance, with respect to nuclear reactions humanity (1) had the implemented capacity of access to radioactivity, (2) this *made available* the safety regulator of controlling chain reactions, (3) determining how to control chain reactions was implemented (through experimentation and calculation), (4) this *unlocked* the capacity to use chain reactions (in a controlled fashion), (5) and the capacity of using chain reactions was implemented.



Limitations and extensions to this method:

- It’s difficult to tell which of the unlocked capacities to implement at a particular step. But we’ll assume some sort of decision process exists for optimizing that.
- Capacities may be good temporarily, but if other capacities are not implemented in time, they may become harmful (see the loss unstable states idea).
- Implementing capacities in this way isn’t necessarily optimal because this approach does not allow for temporary bad effects that yield better results in the long run.
- Capacities do not necessarily stay unlocked forever due to interactions with other capacities that may be implemented in the interim.
- A locked capacity may be net good to implement if a safety regulator is implemented before the downside effects could take place (this is related to handling cluelessness).
- The detailed interaction between capacities and planning which to develop in which order resembles the type of problem the TWEAK planner was built for and it may be one good starting point for further research.
- In more detail, how can one capacity prevent the negative effects of another?

[AN #81]: Universality as a potential solution to conceptual difficulties in intent alignment

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Published a year ago, this sequence of five posts introduced the idea of *ascription universality*. I didn't really get it on a first reading, and only recently read it in enough detail that I think I understand the main ideas. This entire newsletter will focus on ascription universality; treat all of it as a "Highlight".

The key idea of these posts is that of *universality*: when we can say that some agent "knows everything that any other agent could know". Of course, there will always be some agent with arbitrarily complex beliefs, but we could hope to have agents that know everything that is known by any agent of some complexity class (e.g. agents that run in polynomial time). However, there are still simple programs that can have complex *true* beliefs: we could make agents that believe that P != NP and that P = NP, and one of those agents will fall into this class. However, these agents are somewhat degenerate -- they don't have a good reason to *trust* that their belief is true, and so we don't want to worry about them. Since we want to talk about *justified* belief, let's consider a property based on which of the two agents *we* would trust.

So, for now let's adopt an informal definition: a program A[C] is universal with respect to some class of programs C if we would trust any beliefs reported by A[C], no matter what beliefs we hear reported by programs in C. (I'll talk about the specific notion of "ascription universality" later.) Notably, we need to trust A[C] even over our own beliefs, that is, if A[C] believes something, we discard our position and adopt A[C]'s belief.

Note that this is still a very strong criterion. Consider the question of whether AI will lead to x-risk: if one agent thinks about fixed-point theorems, while another studies neural nets, and a third tries to predict the response of democratic institutions, the agent A[C] must compete with *everything* that these agents (justifiably) conclude. For now, let's set aside the question of how to build a universal agent, and instead understand what it enables us to do.

Informed oversight (revisited) *(Paul Christiano)* (summarized by Rohin): Introduced in this older post, we consider a setting in which an overseer is training an agent, and the overseer is smarter than the agent. Intuitively, this should work out well, because if the agent ever tries to do something bad, the overseer should notice this (since it is smarter), and penalize it. However, the problem is that the agent might *in the course of computing an answer* come up with considerations that the overseer can't easily find. For example, an agent tasked with creating novel literature might end up

plagiarizing from one book; the overseer may not be able to search through all the books to detect the plagiarism. However, if the overseer is universal, then it should also know that the agent's output is plagiarized from the book (because the agent knows that), and it can penalize it appropriately.

We can even construct a semi-formal argument (see the post for the math). Let us assume that there is some true utility function U, and when the agent takes some action a, the overseer provides it with reward r = E[U(a)], where the expectation is relative to the overseer's beliefs. Let us assume away problems of distribution shift: in particular, we'll assume that when choosing between actions a1 and a2, the agent always chooses the action with higher reward. Now suppose that the agent chooses action a1. We can conclude that a1 has higher reward than a2. This means that the observer believes that a1 has higher true utility than a2. But by universality, we always trust the observer over the agent and ourselves, and so we must also believe that a1 has higher true utility than a2, which means that we believe the agent is choosing the right action.

Rohin's opinion: This application of universality is the one I believe the most; even intuitively it seems like the entire informed oversight problem is that the overseer *doesn't* know everything that the agent knows, and universality straightforwardly fixes that. So I see this as the motivating example for caring about universality at all.

Worst-case guarantees (Revisited) *(Paul Christiano)* (summarized by Rohin): Let's start considering distribution shift again: we imagine that the agent has already been trained, and there is no overseer any more. When the agent goes off distribution, it may not choose the action that would have gotten the highest reward, and in fact it could be chosen maliciously. As a silly motivating example, perhaps the agent learned the policy "behave well until I see that RSA-2048 (a product of two very large primes) has been factored"; such a policy will behave perfectly during training but will probably commit a treacherous turn eventually.

In this setting, we would like to have a worst-case guarantee, that the model will never behave unacceptably. However, we need to enforce this guarantee at training time, *before* we know what situations will arise at test time. Universality of the overseer isn't enough to guarantee this, because universality only means that the overseer knows what the agent *currently* knows; the agent can't really be said to "know" that it will deceive us in the future.

Nonetheless, it seems like the solution is very related to universality, in that we would like to "peer inside" the agent to figure out how it works, to make sure there aren't situations in which it would behave unacceptably. We could use adversarial training for this purpose: train an adversary to find inputs on which the model behaves unacceptably, and then train the model on those inputs to not do that. This will require interpretability techniques, so that the adversary actually stands a chance of finding the relevant inputs. In order to handle cases like RSA-2048 above, the adversary should also be able to provide "pseudo-inputs" like "the world in which the agent sees that RSA-2048 is factored" (otherwise the adversary would have to factor RSA-2048 itself to demonstrate an input on which the agent would fail, which is far too difficult), while not being allowed to provide pseudo-inputs like "the world in which the agent kills all humans" (which no agent can behave acceptably on). Ultimately, we need to make sure the adversary will always be able to find a bad pseudo-input if one exists, without making it too difficult for the agent to behave acceptably on all pseudo-inputs. See also relaxed adversarial training (AN #70).

Rohin's opinion: I was skeptical of the previous iteration of this post (AN #43), which proposed adversarial training, verification, and interpretability, because it seemed like none of them would allow us to get a *guarantee* on performance. I'm significantly more optimistic after reading this post, mainly because a couple of previous confusions have been resolved:

1. The point of verification is not that we can prove a theorem saying "this agent is beneficial"; the point is that by making *relaxations* (pseudo-inputs), a technique commonly used in formal verification, we can reduce the burden on the other methods being used (such as adversarial training).

2. Similarly, the point of interpretability is not to help *us* understand what the agent is doing or will do, it's to help the *overseer* (or adversary in adversarial training) understand that. Unlike us, the overseer / adversary can scale up along with the agent itself.

I still think that it would be hard to get a guarantee with adversarial training, given that adversarial training has to eliminate *all* vulnerabilities. On the other hand, it only has to find all of the settings where the agent is *maliciously optimizing against us*, which you might hope is a more natural category that is easier to identify without looking too much at particular inputs. This seems like an empirical question on which we'll hopefully get data, though even if it works in all cases that we see, that doesn't rule out the possibility that we failed to notice some issue that will only be triggered in the future (as in the RSA-2048 example).

Universality and model-based RL *(Paul Christiano)* (summarized by Rohin): So far, we've been talking about the model-free setting, where the overseer provides the incentives. What about model-based RL? Here, we might want to learn separate distributions over models and utility functions using iterated amplification or HCH, and then plan using any off-the-shelf algorithm, such as MCTS. The first new problem that arises is that our distribution over models may be malicious (see benign induction). Using universality, we can solve this problem in two different ways. First, we could use HCH to filter out malicious models. By universality, since a malicious model knows what its goals are, so does HCH, and thus HCH will detect it. Second, we could ask HCH to give us a good prior over models. By universality, this prior over models must be better than the prior used by the malicious models, and so the malicious models' predictions will be worse than that of HCH as a whole, and so the malicious models will not be able to affect the computation much.

The second new problem in the model-based setting is that it isn't clear how our utility function can extract all the relevant information from the model. For example, if the model is a detailed simulation of physics, how do you find the "humans" and evaluate how "happy" they are? Universality is at least necessary for this: if there are facts the model "knows", those facts may be relevant to the utility calculation, and so the utility function also needs to know them. We might hope that universality is sufficient: in the previous case, we might think that the physics simulation "knows" whether humans are happy. But this depends on details of what "know" means that haven't yet been fleshed out, and it depends on what "kind" of universality we can achieve in practice.

Universality and consequentialism within HCH *(Paul Christiano)* (summarized by Rohin): In all the settings discussed so far, we hope to use HCH (AN #34). What if HCH itself contains malicious optimization? Specifically, recall that HCH involves an arbitrarily large tree of humans breaking questions down into subquestions. Memetic selection on 10^100 humans should be *much* stronger than the selection done by

evolution to produce us, and so could be quite dangerous. We would like to use HCH to check whether its own computations are malicious.

So, we'll define a pair of mutually recursive functions, Filter and Filtered-HCH. Filter takes as input a transcript from HCH, and using filtered-HCH it finds the best argument suggesting that the transcript is problematic. It then uses filtered-HCH to ask whether, in light of this argument, the transcript should be treated as problematic. This check is repeated thrice for robustness. (Note that the original HCH treats the transcript as trusted computations, while the checker treats it as an untrusted input.) Filtered-HCH has the same type signature as HCH (i.e. it produces answers to questions). Just like HCH, it computes an answer to the question, with the help of recursive calls to filtered-HCH to get answers to subquestions. Instead of returning the answer directly, it calls Filter to check that the computation done was problematic, and returns the special symbol $\perp$ instead of the answer if so.

The basic hope is that if there's some malicious subcomputation implemented "on top of" the base agents in HCH, and it tries to break HCH, then it "knows" why that attempt could work, and so if filtered-HCH is universal with respect to computations that can be instantiated by (smaller versions of) HCH, then filtered-HCH will also "know" that the computation could break HCH, and will censor it as a result. Consider the threat model from Reliability amplification (AN #44), where the chance of any particular HCH computation with budget n becoming malicious is ε, independent of the input or any past HCH calls. Then, intuitively, filtered-HCH (which needs a higher budget, say kn) has a $k\varepsilon$ chance of becoming malicious, but since the check is repeated three times and a majority vote is taken, the chance of a malicious computation getting through is $k^2\varepsilon^2$, which will be much smaller than the original ε if $k^2 \varepsilon << 1$. (This is significantly better than the scheme in Reliability amplification (AN #44), because the question "is there a problem?" is a binary yes/no question, and so majority vote works a lot better.) The harder case is when specific inputs cause very high probabilities of failure (which I believe is the threat model from Security amplification (AN #45)). I particularly don't understand this section, but I believe the hope is that since the malicious computation is implemented "on top of" the (aligned) humans, only a small "part" of the computation can be malicious, and this can be detected by a *small* instance of Filtered-HCH on which the malicious computation cannot be implemented.

Rohin's opinion: I don't understand this post that well, so I'll refrain from opining on it. I think I would benefit from a more concrete example of how malicious optimization could be implemented "on top of" HCH, even when the humans are restricted to a small set of inputs on which they are "secure".

Towards formalizing universality *(Paul Christiano)* (summarized by Rohin): We've now seen a bunch of applications of universality. So what is *ascription* universality? It's a specific formalization of the concept of universality. Recall our informal definition so far: a program A[C] is universal with respect to some class of programs C if we would trust any beliefs reported by A[C], no matter what beliefs we hear reported by programs in C. But how can a program "report beliefs"? For HCH, to find out its belief for some question, we simply ask HCH the question and take its answer as its "belief". However, we can't do this for arbitrary computations: in particular, if our agent is deliberately hiding a belief from us (and so wouldn't report it if asked), we shouldn't simply say that it doesn't have that belief. This is where *ascriptions* come in: an ascription is any way of ascribing beliefs to the agent. Simply asking the agent and trusting its answer is one possible ascription. Another class of ascriptions would look at the agent's code and memory, and make inferences about the agent's beliefs from

that. We want to say that HCH will dominate *any* "reasonable" ascription of beliefs to the agent, where an ascription procedure is "reasonable" if it isn't "doing the hard work"; for example, if an agent always says "Yes", it is unreasonable to ascribe to it the belief that Fermat's Last Theorem is true (even though when given the question "Is Fermat's Last Theorem true" it answers yes).

Putting it all together, A[C] is *ascription universal* with respect to some class of programs C if, for every "reasonable" ascription procedure and program c in C, A[C]'s beliefs (which we find by asking it) *epistemically dominate* the beliefs ascribed to c. A set of beliefs B1 *epistemically dominates* another set of beliefs B2 *from our perspective* if, given arbitrary information about B1 and B2, for every question our best guess at the answer is what we expect B1 would answer.

So far we've been talking about A[C] dominating an entire class of programs C. We might call this *idealized* ascription universality, and hope that it applies to systems with perfect optimization, such as debate (AN #5) under optimal play. In practice, we'll never get perfect optimization, and our overseer isn't going to be able to think of *everything* that any agent could be thinking about. Instead, we could aim for *practical* ascription universality, in which A[c] dominates a particular computation c, and it gets to look at the computation c as a hint. This would allow an overseer to train an agent, while epistemically dominating that agent. Effectively we're asking two questions: first, would our procedure be universal if the optimization was performed perfectly? Second, is weight sharing / interpretability strong enough to solve the optimization problems needed for universality?

The blog post also has some informal thoughts on why we should expect that a careful Bayesian should be able to epistemically dominate computations found via machine learning (which is why we might hope that HCH could be ascription universal).

Rohin's opinion: I was initially a bit confused by the decomposition into the idealized and the practical variant, but I think it makes more sense now: the assumption of *perfect* optimization gives you a lot of mileage. For example, A[C] might choose a particular computation c in C, and then use its beliefs; if that *choice of computation* were already "perfectly optimized" for convincing us, then we would always be at least as convinced of A[C]'s answer as the answer of any other c in C. Debate under optimal play is pretty similar: under optimal play, the agent only tells us about the most important considerations, and only brings up other considerations if its opponent asks about them; you would hope that this is like selecting the best c in C and reporting its reasons for its beliefs.

What's happening here is that the perfect optimization allows us to select the best c in C, or the best argument, etc. "for free" (i.e. without using any computation), giving us huge savings; everything else just has to present the results properly. Consider again the question of whether AI will lead to x-risk: while there are different agents thinking about fixed-point theorems, neural nets, and democratic institutions, perfect optimization lets us "pick" the agent whose arguments are most important. Then, we should only need about as much computation as the agent itself.

However, "presenting the results properly" is non-trivial: A[C] needs to report the beliefs *in plain English*, even if the computation c in C has its beliefs "encoded" (as in the case of a detailed physics simulation that can be said to "know" that humans are happy, but doesn't say that in plain English). It's not clear to me whether this will work out.

You might very reasonably be worried that perfect optimization of "the most convincing argument" will get you something that is convincing for the wrong reasons, and is not actually correlated with the truth. The definition of universality doesn't preclude this: it simply says that we trust A[C]'s beliefs over our own beliefs and any other agent's beliefs; this does hold for convincing but invalid arguments. I believe (but am not sure) that the hope is to deal with this by restricting inputs to the human to a small secure space, and by the techniques discussed above in Universality and consequentialism within HCH.

Nuances with ascription universality *(evhub)* (summarized by Rohin): Consider a formalization of honesty / transparency related to ascription universality: a system is honest, if relative to our expectations, its explicitly stated beliefs give us just as much information as any other way of ascribing beliefs to the system. A system that searches for ways to commit a treacherous turn would not be honest, since it would not state that it was performing this search. However, what if this search was "distilled" into the weights of a neural net (akin to a lookup table)? In that case, since we wouldn't ascribe beliefs to the lookup table, the system could be deemed honest by this criterion. So, what we actually need is for the entire training process to be honest. (A similar critique applies directly to ascription universality.)

Rohin's opinion: This is a good point; it shows that you need your overseer to be ascription universal with respect to the entire training process, not just the final agent.

UML VII: Meta-Learning

(This is the seventh post in a sequence on Machine Learning based on this book. Click here for part I.)

Meta

A more accurate title for this post would be "*Meta, Error bounds, Test Data, Meta-Learning, Error decomposition, and an optional introduction to general probability spaces.*" The first item is me trying to categorize the stuff we've been doing in this sequence so far, which is meta-learning but nonetheless different from the Meta-Learning item, which is about a Machine Learning technique.

Categorizing Machine Learning Insights

Although the book doesn't do this, I think taking some time to reflect on how exactly we've been making progress is very worthwhile. I would categorize the material we've covered so far as follows:

- (1) Carving out relevant and learnable classes out of ML problem space
  - (1.1) Defining such classes and establishing results about their limits
    - **Examples:** Linear Predictors (post 4), Convex (bounded, Lipschitz / smooth) problems (post 5), the negative result on convex problems (post 5)
  - (1.2) Finding algorithms that learn these classes and prove that they work.
    - **Examples:** Perceptron (post 4), Linear Programming (post 4), applied linear algebra (post 4), Stochastic Gradient Descent (post 6)
- (2) Extending the usability of such classes
    - **Examples:** Primarily surrogate loss functions (post 5); also the mapping of inhomogeneous linear problems to linear problems (post 4), even if it's just a detail. There will be more stuff here in future posts.
- (3) General (as in, widely applicable) theoretical work
    - **Examples:** The general learning model, work on overfitting, the PAC learning framework, the basic error decomposition, the ERM approach (all post 1), the uniform convergence property, the No-Free-Lunch Theorem, the VC dimension, the fundamental theorem of statistical learning (all post 2), the nonuniform learning framework, the SRM approach, the minimal description length, the basics of computational complexity (all post 3), meta-learning, error bounds via test data, and advanced error decomposition (later in this post)

Not quite covered: coming up with learning algorithms. So far, that's just Stochastic Gradient Descent, which I have listed as a solution to solve convex Lipschitz/smooth bounded problems, but it's not restricted to that. Also, neural networks supposedly have an impressive track record of fitting all sorts of functions, and I don't expect there to be analogous theoretical results.

Based on this categorization (which is not perfect – much of (3) is only about binary classification), the entire first part of the book which was covered in posts I-III is only

about broad theoretical work, hence "foundations". The remaining book jumps around between all three. Note that my adaption isn't always linear.

Error Bounds

One way to motivate this chapter is that we *wish to find better guarantees for the quality of predictors.* We usually either care about the term $\ell(A_{ERM}(S)) - \ell(h^*)$ (where h^* is the predictor with minimal real error in H) or about $\max_{h \in H}[\ell(h) - \ell_S(h)]$. It turns out that both are closely related. We'll call either of these the "gap". Now one can alternatively ask

- "given that I have m data points, what is the smallest upper-bound on the gap that I can prove?"; or
- "given that I want to establish an upper bound of ϵ on the gap, how many data points do I need?"

where, in both cases, the guarantee will hold with some probability $1 - \delta$, rather than with certainty. The first would be an error bound, the second a sample complexity bound, but they're equivalent: given an error bound, one can take the corresponding equation, which will have the form $\ell(A_{ERM}(S)) - \ell(h^*) \leq \langle$some term which depends on m$\rangle$, set the left part to ϵ and solve it for m, which will yield a sample complexity bound. Going into the opposite direction is no harder.

Musings on past approaches

The punchline of this section is that all bounds we have looked at so far are based on the performance of the classifier *on the data it's been trained on*. Given that the output predictor may be highly dependent on patterns that only exist in the training data but not in the real world, this is difficult – and, as we know from the No-Free-Lunch Theorem, even impossible without making further assumptions.

This implies that we must have made such assumptions, and indeed we have. In addition to the empirical error, our current bounds have been based on

- (1) the hypothesis class
- (2) the properties of the classifier that the learner guarantees.

By (2), I mean that the classifier is in the set $\text{argmin}_{h \in H}[\ell_S(h)]$ in the case of A_{ERM}, in the class $\text{argmin}_{h \in H}[\ell_S(h) + \epsilon_n(\delta, h)]$ in the case of A_{SRM}; and in the case of A_{SGD} we've used the update rule in the proof that derived the error bound.

Deriving error bounds with test data

By testing the classifier on data that the learner didn't have access to, we can toss out (1) and (2) in favor of a purely statistical argument, which results in a bound that is more general, much easier to prove, and stronger. The downside is that it requires additional data, whereas the previous bounds were "free" in that regard.

To be more precise, the idea now is that we split our existing data into two sequences S and T; we train a learner on S and we test the output predictor $A(S)$ on T to obtain an estimate of the real error. In symbols, we use $\ell_T(A(S))$ as our estimate for $\ell(A(S))$, where ℓ_T is defined just like ℓ_S only, of course, using T instead of S as the set (this definition is implicit if one simply regards the definition of ℓ_S as applying to ℓ with any set S as a subscript). So $\ell_T(h) = \frac{1}{|T|}|\{(x, y) \in T \mid h(x) \neq y\}|$. We will call the output of ℓ_T the *test error*.

In general, the real error does not equal the test error (in symbols, $\ell(A(S)) \neq \ell_T(A(S))$ because the test error is based on the test sequence T which may be unrepresentative. However, the real error equals the *expected test error* (we'll prove this in the optional chapter at the end of the post). Therefore, we only need an upper-bound on the difference between the expected value and the actual value of random variables (this is the aforementioned statistical argument), and a theorem to that end has already been established by relevant literature: *Hoeffding's Inequality*.

> Let $\theta_1, \ldots, \theta_m$ be i.i.d. RVs that range over the interval $[0, 1]$. Then, for all $\epsilon \in \mathbb{R}_+$, it holds that $\Pr(|\bar{\theta} - E[\bar{\theta}]| > \epsilon) \leq 2e^{-2\epsilon^2 m}$, where $\bar{\theta} := \frac{1}{m}\sum_{i=1}^{m} \theta_i$.

This uses the same notational shortcut that is used all the time for Random Variables (RVs), namely pretending as if RVs are *numbers* when in reality they are *functions*. Importantly, note that $\bar{\theta}$ depends on the input randomness, whereas $E[\bar{\theta}]$ is a constant.

In our case, the random variables we are concerned with are the *point-based loss functions applied to our predictor* for our instances in T, i.e., if $T = ((x_1, y_1), \ldots, (x_t, y_t))$, then our random variables are $\theta_1 = \ell_{(x_1,y_1)}(h), \ldots, \theta_t = \ell_{(x_t,y_t)}(h)$. They are random variables if we regard the points they're based on as unknown (otherwise they'd just be numbers). Their mean, $\bar{\theta} = \frac{1}{|T|}\sum_{i=1}^{|T|} \ell_{(x_i,y_i)}(h)$ equals the test error, $\ell_T(h)$, whose randomness ranges over the choice of all of T. So $\bar{\theta} = \ell_T(h)$ and $E(\bar{\theta}) = \ell(h)$. Given this,

Hoeffding's inequality tells us that $\Pr(|\ell(h) - \ell_T(h)| > \epsilon) \leq 2e^{-2\epsilon^2 m}$, provided that all loss functions range over $[0, 1]$ (if not, then as long as they are still bounded, a generalized version of this theorem exists). Furthermore, if one drops the $|\ |$, the failure probability halves, i.e. $\Pr(\ell(h) - \ell_T(h) > \epsilon) \leq e^{-2\epsilon^2 m}$. Setting $\delta := e^{-2\epsilon^2 m}$ and solving for ϵ, we obtain $\epsilon = \sqrt{\frac{\ln(1/\delta)}{2m}}$. Thus, we have established the following:

> If T is a sequence of i.i.d examples sampled by the same distribution that generated S, but is disjoint from S, then the inequality $\ell_T(A(S)) - \ell(A(S)) \leq \sqrt{\frac{\ln(1/\delta)}{2|T|}}$
>
> holds with probability at least $1 - \delta$ over the choice of T.

For a brief comparison: a bound provided by the (quantitative version of) the fundamental theorem of statistical learning states that $|\ell_S(h) - \ell(h)| \leq \sqrt{C\frac{d + \ln(1/\delta)}{m}}$, provided that H has VC-dimension $d \in \mathbb{N}$. So this bound relies on a property of the hypothesis class (and if that property doesn't hold, it doesn't tell us anything), whereas our new bound always applies. And as a bonus, the new bound doesn't have the constant C but instead has a nice 2 in the denominator.

Meta-Learning

The book now goes on to talk about how this idea can be used for ***model selection.*** Let's begin by summarizing the argument.

Model Selection ...

Suppose we have several possible models for our learning task, or alternatively one model with different parameters (maximal polynomial degree, step size of stochastic gradient descent, etc). Then we can proceed as follows: we split our training data into three sequences, the training sequence S, the *validation sequence* V, and the test sequence T. For each model we look at, we train the corresponding learner A on S, then we test all output hypotheses $A_1(S), \ldots, A_k(S)$ on V, i.e. we compute $\ell_V(A_1(S)), \ldots, \ell_V(A_k(S))$, and pick one with minimal validation error, i.e. we pick $A_j(S)$

such that $j \in \operatorname{argmin}_{i \in \{1,\dots,k\}} \ell_V(A_i(S))$. Finally, we compute $\ell_T(A_j(S))$ and use that value to argue that the predictor performs well (but not to change it in any way).

Since the learners $A_1(S), \dots, A_k(S)$ haven't been given access to V, the validation error will be unbiased feedback as to which learner performs best, and since even *that* learner has been computed without using T, the test error is yet again unbiased feedback on the performance of our eventual pick $A_j(S)$. That is, provided that there are no spurious correlations across our data sets, i.e. provided that the i.i.d. assumption holds.

One can be even more clever, and trade some additional runtime for more training data. The test sequence T remains as-is, but rather than partitioning the remaining data in S and V we partition it into a bunch of blocks of equal size, say k blocks, and then, for each such block i, we

- train our learner A on all blocks $S_{\neq i}$
- compute the error on block i, i.e. $\ell_{S_i}(A(S_{\neq i}))$

This way, we obtain k different validation errors (one for each classifier which was trained on nine blocks and tested on the tenth); now we take the mean of all of them as our ultimate validation error. We do this entire thing for each learner A, take the one that performed best, and then we can even retrain that learner on all k blocks to make sure we fully utilize all data we have.

Finally, just as before, we test that final predictor using the test data and output the result. This entire process is called k***-fold cross validation.***

... is just meta-learning

While this is all well and good, I think it's important to understand that the specifics are non-fundamental. What is novel here is the concept of obtaining unbiased feedback by computing the empirical error of a predictor based on a sequence that has not been used to learn that predictor. Beyond that, the observation is that it might be useful to do meta-learning, i.e. have several levels of learning, i.e. partition all possible hypotheses in a bunch of different classes, take the best from each class, and then take the best across all classes.

But the notion of exactly two levels is arbitrary. Suppose we have a bunch of models, and each model has some parameters. Then we can partition our training data in *four* sequences S, V_1, V_2, T. For each model, we choose a bunch of plausible parameter

values, so say we have n models and each one has m parameter settings. For each such setting, we train the learner on S and test it on V_1. We pick the optimal one among those m; this will be the predictor with the optimal parameters for our model. We do this for all n models, leading to n predictors, all of whom use the optimal parameters for their model. We test these n predictors using V_2 and output the optimum yet again. Finally, we test that predictor on T.

Similarly, one could imagine four levels, or however many. The crucial thing is just that choosing the optimal predictor of the next lower level has to be done with data that hasn't been used to learn that predictor. So in the above example, for each [model and particular parameter setting], the respective output hypothesis just depends on S. If we then compute the optimal predictor of that model by looking at which of the m predictors (corresponding to the m parameter settings) performed best on V_1, this output predictor also depends on V_1. In fact, the procedure I just described is just one kind of learning algorithm that uses both S and V_1 as training data. Similarly, if we take the predictor with the smallest error on V_2 across all n models to obtain our final predictor, then this predictor depends on S and V_1 and V_2 – in fact, the procedure I just described is just one kind of learning algorithm that uses S and V_1 and V_2.

Differently put, to derive a predictor for a particular problem, one might choose to apply meta-learning, i.e. choose k different models, train a predictor in each model, and then select the best of those k predictors, based on their performance on training data which hasn't been used to train them – and each of those k smaller problems is itself just another problem, so one might again choose to apply meta-learning, and so on.

The trick from the last section for saving data can also be used with more than 2 levels, although it quickly increases runtime.

The idea of using a test sequence is more fundamental (i.e. not just another level), since it's only used to obtain a certificate of the quality of the final predictor, but *not* for learning (in any way). As soon as one violates this rule (for example by choosing another predictor after the first one showed poor performance on the test data), the bound from Hoeffding's inequality is no longer applicable, since the test error is no longer an unbiased estimate of the real error.

Error Decomposition

So far, we've decomposed the error of a predictor like so:

$$\ell(h) = (\ell(h) - \ell(h^*)) + \ell(h^*)$$

where $h^* \in \operatorname{argmin}_{h \in H}[\ell(h)]$. The term $\ell(h^*)$ is the *approximation error*, which is the lowest error achievable by any predictor in our hypothesis class ("how well does this class approximate the problem?") and the term $(\ell(h) - \ell(h^*))$ is the *estimation error* ("how well did we estimate the best predictor?").

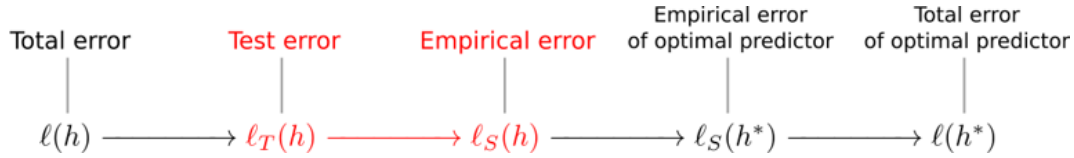
The usefulness of this approach is limited. One doesn't actually know the approximation error – if learning fails, it's not clear which of the two errors is at fault. Furthermore, both overfitting and underfitting is part of the estimation error, which suggests that it should be divided up further (whereas the approximation error can stay as-is). This is what we do now. Our new decomposition will rely heavily on having the kind of unbiased feedback from some independent data which we've been talking about in the previous chapter. Note, however, that while I'll only talk about the test sequence T going forward, everything applies equally if one uses a validation sequence instead (i.e. training data which has not been given to the learner). In fact, using such data is the way to go if one intends to change the learner in any way upon seeing the various parts of the decomposition.

I find the notation with differences not very illustrative, which is why I'm making up a different one. The above decomposition can be alternatively written as

$$\ell(h) \longrightarrow \ell(h^*)$$

where we decompose the left term into the value of the arrow plus that of the right term. (Each error corresponds to adding "+[right term − left term]". If there are more than two terms, we decompose the leftmost term into the sum of all errors plus the rightmost term (actually, any term in the chain equals the sum of all errors to its right plus the rightmost term).

That said, here is a more sophisticated error decomposition:



The red terms are the ones we have access to. Note that it is not the case that each step is necessarily negative, so the values are not monotonically decreasing. Now let's look at this more closely.

"$\ell(h) \longrightarrow \ell_T(h)$" can be bounded using Hoeffding's inequality.

Let's pause already and reflect on that result. Our only terminal value is to maximize the leftmost term; the first arrow can be bounded tightly; and the second term is

known. It follows that, if the test error is small we can stop right there (because the real error is probably also small). In fact, that was the point of the previous chapter. The remaining decomposition is only necessary if the test error is larger than what is acceptable.

"$\ell_T(h) \longrightarrow \ell_S(h)$" is the difference between the unbiased and the biased error estimate – the error on the data the learner had no access to minus the error on the data it did have access to. It measures how much we overfit – which is quite useful given that this is another term we have access to. Thus, if the test error is large but the empirical error small, we might want to change our learner such that fits the training data less closely. However, we don't have any guarantee in this case that overfitting is the only problem. It's possible that we overfit *and* the approximation error is large.

If the test error and the empirical error are both large, that's when we're interested in the remaining part of this decomposition.

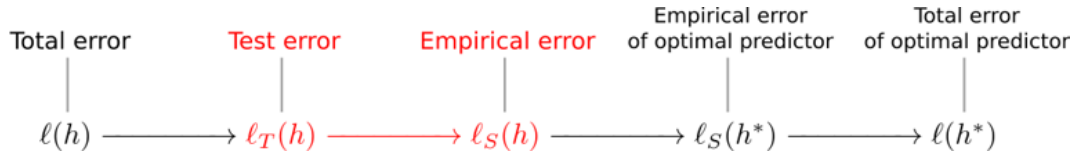
"$\ell_S(h^*) \longrightarrow \ell(h^*)$" is usually negative, at least if we're using A_{ERM} or something close to it.

"$\ell_S(h^*) \longrightarrow \ell(h^*)$" can also be bounded by Hoeffding's inequality, because h* does not depend on S (unless we've messed with the hypothesis class based on the training data).

Finally, $\ell(h^*)$ is the approximation error from our previous decomposition. It doesn't make sense to decompose this further – if the approximation error is large, our hypothesis class doesn't contain any good predictors and learning it is hopeless.

Thus, based on the previous 3 paragraphs, the empirical error is unlikely to be significantly larger than the total error of the output predictor, i.e. the approximation error. (It might be significantly smaller, though.) It follows that if the empirical error is high, then so is the approximation error (but not vice-versa), *with the important caveat that one needs to be confident in the learner's ability to minimize empirical risk.* If the problem is learning starcraft, then it's quite plausible that minimizing the empirical risk is so difficult that the empirical risk could be large even if the approximation error is very small.

To summarize...



... one can reason like this:

- Test error...
  - is small → real error is probably small too, so everything's good
  - is large → empirical error...

- is small → the learner probably suffers from overfitting[1]
- is large → the empirical error seems...
  - easy to minimize → approximation error is probably large
  - hard to minimize → ??? (the problem is just hard)

If the approximation error is large, then the problem is unrelated to our learner. In that case, the hypothesis class H may be "too small"; perhaps it contains only linear predictors and the real function just isn't linear. Alternatively, it could also be the case that the real function is approximately linear, but we just didn't choose features that represent meaningful properties of the data points. For example, if the problem is learning a function that classifies emails as spam / not spam, one could imagine that someone was just really wrong about which features are indicative of spam emails. This is the earliest point of failure; in this case, learning might be impossible (or at least extremely difficult) even with a good hypothesis class and a good learner.

[1] While this is technically true, one might actually be fine with some amount of overfitting. If so, then this case isn't quite satisfactory, either; in particular, one doesn't know whether the approximation error is large or not. Another way to approach the question is to train the learner on parts of the training sequence first, say $S_1 \subsetneq \cdots \subsetneq S_k = S$, and examine the test error for each learner $A(S_j)$. If it goes down as we use more data, that's some evidence that the approximation error is low.

In general, this chapter is far from a full treatment of error decomposition.

Proof that E(test error) = real error

As mentioned, this chapter is completely optional (also note that the post up to this point is roughly as long as previous posts in this sequence). It's long bothered me that I didn't understand how a general probability space was defined, so if you share that frustration, this chapter will provide an answer. However, the connection to Machine Learning is very loose; one needs general probability spaces to prove the result, but the actual proof is quite easy, so this will be a lot of theory with a fairly brief application.

Preamble: General Probability Spaces

Before the general case is introduced, one is generally being taught two different important special cases of probability spaces. The first is the *discrete* one. Here, a probability space is given by a pair (Ω, p), where Ω is a countable sample space and $p : \Omega \to [0, 1]$ a function that assigns to each *elementary event* $\omega \in \Omega$ a probability. For any subset $A \subseteq \Omega$, the probability $P(A)$ is then simply defined as $\sum_{\omega \in A} p(\omega)$. And for a random variable $X : \Omega \to R$, we have the definition

$$(1): E(X) := \int_{-\infty}^{\infty} xf(x)\,dx.$$

This is super intuitive, but it can't model how an event happens at a random point throughout a day. For that, we need the *absolutely continuous* case. This is where no one point has nonzero probability, but intervals do. In other words, one is in the absolutely continuous case whenever the probability distribution admits of a *probability density function* such that, for each interval, the probability of that interval is equal to the integral of that function over that interval. If $\Omega = \mathbb{R}$ (covering only that case here), we can write the probability space as (Ω, f) where for any interval $[a, c] \subseteq \Omega$, we have $P([a, c]) = \int_a^c f(x)\,dx$. This is also rather intuitive – one can just draw an analogy between probability mass and area. A single point has an "infinitely thin" area below it, so it has zero probability mass (and in fact the same is true for any countable collection of points because even the tiniest interval is uncountable), but an interval does have real area below it. The height of the area is then determined by f.

Given a random variable $X : \Omega \to \mathbb{R}$, we have

$$(2): E(X) := \int_{-\infty}^{\infty} xf(x)\,dx.$$

But neither of these is the general case, because one assumes that every point has nonzero probability and the other assumes that no point does. In the general case, a probability space looks like this:

$$(\Omega\ ,\ \Sigma\ ,\ P\ )$$

The first element Ω is the same as before – the sample space. But the Σ is something new. It is called a σ-*algebra,* which is a *set of subsets* of Ω, so $\Sigma \subseteq P(\Omega)$, and it *consists of all those subsets which have a probability*. It needs to fulfill a bunch of properties like being closed under complement and like $\Omega \in \Sigma$, but we don't need to concern ourselves with those.

The probability distribution is a function $P : \Sigma \to [0, 1]$. So it's a function that assigns subsets of Ω a probability, but only those subsets that we declared to have a probability, i.e. only those subsets that are elements of our σ-algebra Σ. It must have a bunch of reasonable properties such as $P(\Omega) = 1$ and $P(A) + P(B) = P(A \sqcup B)$, where the symbol $\sqcup$ is meant to indicate that A and B are disjoint.

(In the discrete case, we don't need to specify Σ because we'd simply have $\Sigma = P(\Omega)$, i.e. every subset has a probability. I'm not sure how it would look in the continuous case, but maybe something like the set all of all subsets that can be written as a countable union of intervals plus a countable union of single points.)

The question relevant for us is now this: given a general probability space (Ω, Σ, P) and a random variable $X : \Omega \to P$, how do we compute the expected value of X? And the answer is this:

$$(3): E(X) := \int_\Omega X(\omega)\ dP(\omega)$$

where $\int$ is the *Lebesgue integral* rather than the Riemann integral. But how does one compute a Lebesgue integral? Fortunately, we don't require the full answer to this question, because some of the easy special cases will suffice for our purposes.

Computing Lebesgue Integrals of Simple Functions

The setting to compute a Lebesgue integral is a bit more general than that of a general probability space, but since we're only trying to do the easy cases anyway, we'll assume our space has the same structure as before. However, we will rename our probability distribution P into μ and call it a *measure,* so that our space now looks like this:

$$(\Omega ,\ \Sigma ,\ \mu)$$

And we'll also rename our random variable into f to make it look more like a general function. So $f : \Omega \to R$ is the function whose Lebesgue integral $\int_X f\ d\mu$ we wish to compute.

Recall that P assigns a probability to the subsets of Ω that appear in Σ, and that the entire space has probability 1. Analogously, μ assigns a *measure* to subsets of Ω, and the entire space has measure 1. So let Y be a subset that is "half" of Ω, then $\mu(Y) = \frac{1}{2}$.

One now proceeds differently than with Riemann integrals. Rather than presenting the general case right away, we begin with the simplest kind of function, namely *indicator functions*. That is a function of the form 1_S for some $S \in \Sigma$, and it is simply 1 on S and 0 everywhere else. And now, even though we don't understand how Lebesgue integrals work in general, in this particular case the answer is obvious:

$$\int_\Omega 1_S \, d\mu := \mu(S).$$

The function 1_S just says "I count S once and I count nothing else", so the integral has to just be one times the size, i.e. the *measure,* of S. Which is μ(S).

So for example, $\int_\Omega 1_Y \, d\mu = \mu(Y) = \frac{1}{2}$ and $\int_\Omega 1_\Omega \, d\mu = \mu(\Omega) = 1$.

Next, if f is not itself an indicator function, but it can be written as a finite weighted sum of indicator functions, i.e. $f = \sum_{i=1}^{m} a_i 1_{S_i}$, then f is called a *simple function* and we define

$$\int_X f \, d\mu := \sum_{i=1}^{m} a_i \mu(S_i)$$

So for example, if $f = 2 \cdot 1_Y$, i.e. the function that is 2 on one half of the space and 0 on the other, then $\int_X f \, d\mu = 2\mu(Y) = 2 \cdot \frac{1}{2} = 1$.

And this where we stop because integrating simple functions is all we need.

An example

Let's return to our setting of general probability spaces. Here, the probability distribution becomes the measure. So let's say our space is $\Omega = [0, \frac{1}{2}] \cup \{10\}$. Our σ-algebra Σ includes all sub-intervals of $[0, \frac{1}{2}]$ (closed or open or half-open), and all such intervals plus the point 10. Our probability distribution P says that $P(\{10\}) = \frac{1}{2}$ and the probability mass of an interval in $[0, \frac{1}{2}]$ is just the size of the interval, so if $[a, c] \subseteq [0, \frac{1}{2}]$, then $P([a, c]) = c - a$. Then, $P(\Omega) = 1$. Recall that P has to assign each set $M \in \Sigma$ a probability, which it now does.

Let's say $X : \Sigma \to R$ is a random variable given by $X = 5 \cdot 1_{[0, \frac{1}{3}]} + 7 \cdot 1_{\{10\}}$. Then, X is a simple function – it is the weighted sum of two indicator functions – and we can compute the expected value of X as

$E(X) = \int_\Omega X \, dP = 5 \cdot P([0, \frac{1}{3}]) + 7 \cdot P(\{10\}) = 5 \cdot \frac{1}{3} + 7 \cdot \frac{1}{2} = \frac{5}{3} + \frac{7}{2} = \frac{31}{6}$.

The actual proof

The real error of h := A(S) is

$$\ell\,(\,h\,)\;=\;D\,(\,\{\,(\,x\,,\;y\,)\;\in\;X\;\times\;Y\;\;\mid\;\;h\,(\,x\,)\;\neq\;y\,\}\,)$$

The test error $\ell_T(h)$ depends on the test sequence. Let m be the length it the sequence, so that $T = ((x_1, y_1), \ldots, (x_m, y_m))$; then to compute the *expected* test error, we have to compute the Lebesgue integral

$$\int_\Omega \ell_T(h)\, dD^m \text{ where } \Omega = (X \times Y)^m$$

So our measure for the space Ω is now D^m. We wish to write $\ell_T(h)$ as a sum of indicator functions. Recall that $\ell_T(h) = \frac{1}{m}|\{(x, y) \in T \mid h(x) \neq y\}|$, i.e. $\ell_T(h)$ counts the number examples which h got wrong. Thus, by defining the sets $W_k := \{((x_1, y_1), \ldots, (x_m, y_m)) \in \Omega \mid h(x_k) \neq y_k\}$ for all $k \in \{1, \ldots, m\}$, we get the equality $\ell_T(h) = \sum_{k=1}^{m} \frac{1}{m} \cdot 1_{W_k}$.

Thus, the integral simply equals the probability mass of the respective sets (multiplied by their respective weights). In symbols,

$$\int_\Omega \ell_T(h)\, dD^m = \sum_{k=1}^{m} \frac{1}{m} D^m(W_k)$$

The i.i.d. assumption tells us that $D^m(W_k) = D(\{(x, y) \in X \mid h(x) \neq y\}) = \ell(h)$ (in fact, this or something similar is probably how it should be formally defined), and therefore the above sum equals $\sum_{k=1}^{m} \frac{1}{m} \ell(h) = \ell(h)$.

Algorithms vs Compute

Two scenarios:

- I take a vision or language model which was cutting edge in 2000, and run it with a similar amount of compute/data to what's typically used today.
- I take a modern vision or language model, calculate how much money it costs to train, estimate the amount of compute I could have bought for that much money in 2000, then train it with that much compute

In both cases, assume that the number of parameters is scaled to available compute as needed (if possible), and we generally adjust the code to reflect scalability requirements (while keeping the algorithm itself the same).

Which of the two would perform better?

CLARIFICATION: my goal here is to compare the relative importance of insights vs compute. "More compute is actually really important" is itself an insight, which is why the modern-algorithm scenario talks about compute cost, rather than amount of compute actually used in 2000. Likewise, for the 2000-algorithm scenario, it's important that the model only leverage insights which were already known in 2000.

What is Life in an Immoral Maze?

Previously in sequence: Moloch Hasn't Won, Perfect Competition, Imperfect Competition, Does Big Business Hate Your Family?

This post attempts to give a gears-level explanation of maze life as experienced by a middle manager in systems with many levels of management, as depicted in *Moral Mazes.*

The 'maze level' of corporations differs wildly. These dynamics do not reliably fully take over until you have many levels of management. Questions of what causes high maze levels will be dealt with in future sections.

Again, if you have not yet done so, you are highly encouraged to read or review Quotes from Moral Mazes. I will not have the space here to even gloss over many important aspects.

An Immoral Maze can be modeled as a super-perfectly competitive job market for management material. All the principles of super-perfect competition are in play. The normal barriers to such competition have been stripped away. Too many 'qualified' managers compete for too few positions.

If an aspirant who does not devote everything they have, and visibly sacrifice all slack, towards success, they automatically fail. Those who do make such sacrifices mostly fail anyway, but some of them "succeed". We'll see later what success has in store for them.

The Lifestyle of a Middle Manager

> *At the managerial and professional levels, the road between work and life is usually open because it is difficult to refuse to use one's influence, patronage, or power on behalf of another regular member of one's social coterie. It therefore becomes important to choose one's social colleagues with some care and, of course, know how to drop them should they fall out of organizational favor. (Moral Mazes, Location 884, Quote #117)*

We have this idea that there is work and there is not-work, and once one leaves work one is engaged in not-work distinct from work. We also have this idea that there are things that are off limits even at work, like sexual harassment.

For a person without anyone reporting to them, who is 'on the line' in the book's parlance, this can be sustained.

For those in middle management who want to succeed, that's not how things work. Everything you are is on the table. You'd better be all-in.

You will increasingly choose your friends to help you win. You will increasingly choose your hobbies, and what you eat, and your politics, and your house, and your church, and your spouse and how many kids you have, to help you win. And of course, you will choose your (lack of) morality.

In the end, you will sacrifice *everything,* and I mean everything, that you value, in any sense, to win.

If the job requires you to move, anywhere in the world, you'll do it, dragging your nuclear family along and forcing all of you to leave behind everything and everyone you know. Otherwise, you're just not serious about success.

Slack will *definitely* not be a thing.

Your time is especially vulnerable.

> *Higher-level managers in all the corporations I studied commonly spend twelve to fourteen hours a day at the office. (Location 1156, Quote #120, Moral Mazes)*

This is the result of total competition between producers – the managers are effectively rival producers trying to sell themselves as the product.

The market for managers is seen, by those who make the decisions, as highly efficient.

If managers were seen as wildly different in terms of talent, intelligence, or some other ability that helped get things done, that would help a lot. You could afford to be a little quirky, to hold on to the things you value most, without losing the game entirely. Your success will be *influenced* by your personality and dedication, but nothing like *solely determined* by them.

Alas, the perception in these mazes is exactly the opposite.

> *See, once you are at a certain level of experience, the difference between a vice-president, an executive vice-president, and a general manager is negligible. It has relatively little to do with ability as such. People are all good at that level. They wouldn't be there without that ability. So it has little to do with ability or with business experience and so on. All have similar levels of ability, drive, competence, and so on. What happens is that people perceive in others what they like—operating styles, lifestyles, personalities, ability to get along. Now these are all very subjective judgments. And what happens is that if a person in authority sees someone else's guy as less competent than his own guy, well, he'll always perceive him that way. And he'll always pick—as a result—his own guy when the chance to do so comes up. (Location 1013, Quote #87, Moral Mazes)*

It is known that most people 'don't have what it takes' to be a manager. This is clearly true on many levels. Only one of them is a willingness to fully get with the program.

Once you get several levels up, the default assumption is that everyone is smart enough, and competent enough. That the object-level is a fully level playing field. The idea that someone can just *be better at doing the actual job* doesn't parse for them.

All remaining differences are about negative selection, about how hard you want it and are willing to sacrifice everything, or about how well you play political games. Nor do they much care whether you succeed at your job, anyway.

Some additional supporting quotes on that follow. A large portion of the quotes reinforce this perspective.

If you can't work smart, work hard:

> *When asked who gets ahead, an executive vice-president at Weft Corporation says: The guys who want it [get ahead]. The guys who work. You can spot it in the first six months. They work hard, they come to work earlier, they leave later. They have suggestions at meetings. They come into a business and the business picks right up. They don't go on coffee breaks down here [in the basement]. You see the parade of people going back and forth down here? There's no reason for that. I never did that. If you need coffee, you can have it at your desk. Some people put in time and some people work. (Location 992, Quote 29, Moral Mazes)*

But everyone at this level works hard, which was more about showing you work hard than the results of the work, because concrete outcomes don't much matter:

> *As one manager says: "Personality spells success or failure, not what you do on the field." (Location 1383, Quote 33, Moral Mazes)*

It's not like there were ever objective criteria:

> *Managers rarely speak of objective criteria for achieving success because once certain crucial points in one's career are passed, success and failure seem to have little to do with one's accomplishments. (Location 917, Quote 42, Moral Mazes)*

Which makes sense, because if everyone is the same, then concrete outcomes are just luck:

> *Assuming a basic level of corporate resources and managerial know-how, real economic outcome is seen to depend on factors largely beyond organizational or personal control. (Location 1592, Quote 46, Moral Mazes)*

I am supremely confident that this perspective is completely bonkers. There is huge differential between better and worse no matter how high up you go or how extreme your filters have already been. But *what matters here is what the managers believe.* Not what is true. Talent or brilliance won't save you if no one believes it can exist. If noticed it will only backfire:

> *Striking, distinctive characteristics of any sort, in fact, are dangerous in the corporate world. One of the most damaging things, for instance, that can be said about a manager is that he is brilliant. This almost invariably signals a judgment that the person has publicly asserted his intelligence and is perceived as a threat to others. What good is a wizard who makes his colleagues and his customers uncomfortable? (Location 1173, Quote 88, Moral Mazes)*

How do things get so bad?

That's the question we'll look at an aspect of next post. From here I anticipate 3-5 day gaps between posts.

Questions that will be considered later, worth thinking about now, include: How does this persist? If things are so bad, why aren't things *way worse?* Why haven't these corporations fallen apart or been competed out of business? Given they haven't, why hasn't the entire economy collapsed? Why do regular people, aspirant managers and otherwise, still think of these manager positions as the 'good jobs' as opposed to picking up pitchforks and torches?