Best of LessWrong: January 2019

1. Building up to an Internal Family Systems model
2. The 3 Books Technique for Learning a New Skilll
3. Some Thoughts on My Psychiatry Practice
4. Book Summary: Consciousness and the Brain
5. Disentangling arguments for the importance of AI safety
6. What are the open problems in Human Rationality?
7. Announcement: AI alignment prize round 4 winners
8. Strategy is the Deconfusion of Action
9. Combat vs Nurture & Meta-Contrarianism
10. S-Curves for Trend Forecasting
11. Megaproject management
12. Less Competition, More Meritocracy?
13. From Personal to Prison Gangs: Enforcing Prosocial Behavior
14. Reframing Superintelligence: Comprehensive AI Services as General Intelligence
15. Book Review: The Structure Of Scientific Revolutions
16. Alignment Newsletter #39
17. Sequence introduction: non-agent and multiagent models of mind
18. Visualizing the power of multiple step selection processes in JS: Galton's bean machine
19. Book Recommendations: An Everyone Culture and Moral Mazes
20. CDT=EDT=UDT
21. Supervising strong learners by amplifying weak experts
22. Does anti-malaria charity destroy the local anti-malaria industry?
23. Alignment Newsletter #40
24. Anthropic probabilities: answering different questions
25. Learning with catastrophes
26. What are good ML/AI related prediction / calibration questions for 2019?
27. "The Unbiased Map"
28. Buy shares in a megaproject
29. Future directions for narrow value learning
30. How could shares in a megaproject return value to shareholders?
31. Two More Decision Theory Problems for Humans
32. No surjection onto function space for manifold X
33. On Abstract Systems
34. What is narrow value learning?
35. A general framework for evaluating aging research. Part 1: reasoning with Longevity Escape Velocity
36. Failures of UDT-AIXI, Part 1: Improper Randomizing
37. Optimizing for Stories (vs Optimizing Reality)
38. How much can value learning be disentangled?
39. [Speech] Worlds That Never Were
40. What is a reasonable outside view for the fate of social movements?
41. Which approach is most promising for aligned AGI?
42. Is Agent Simulates Predictor a "fair" problem?
43. What math do i need for data analysis?
44. What are the components of intellectual honesty?
45. One Website To Rule Them All?
46. Comments on CAIS
47. Why don't people use formal methods?
48. Littlewood's Law and the Global Media
49. "Forecasting Transformative AI: An Expert Survey", Gruetzemacher et al 2019
50. Prediction Contest 2018: Scores and Retrospective

Best of LessWrong: January 2019

1. Building up to an Internal Family Systems model
2. The 3 Books Technique for Learning a New Skilll
3. Some Thoughts on My Psychiatry Practice
4. Book Summary: Consciousness and the Brain
5. Disentangling arguments for the importance of AI safety
6. What are the open problems in Human Rationality?
7. Announcement: AI alignment prize round 4 winners
8. Strategy is the Deconfusion of Action
9. Combat vs Nurture & Meta-Contrarianism
10. S-Curves for Trend Forecasting
11. Megaproject management
12. Less Competition, More Meritocracy?
13. From Personal to Prison Gangs: Enforcing Prosocial Behavior
14. Reframing Superintelligence: Comprehensive AI Services as General Intelligence
15. Book Review: The Structure Of Scientific Revolutions
16. Alignment Newsletter #39
17. Sequence introduction: non-agent and multiagent models of mind
18. Visualizing the power of multiple step selection processes in JS: Galton's bean machine
19. Book Recommendations: An Everyone Culture and Moral Mazes
20. CDT=EDT=UDT
21. Supervising strong learners by amplifying weak experts
22. Does anti-malaria charity destroy the local anti-malaria industry?
23. Alignment Newsletter #40
24. Anthropic probabilities: answering different questions
25. Learning with catastrophes
26. What are good ML/AI related prediction / calibration questions for 2019?
27. "The Unbiased Map"
28. Buy shares in a megaproject
29. Future directions for narrow value learning
30. How could shares in a megaproject return value to shareholders?
31. Two More Decision Theory Problems for Humans
32. No surjection onto function space for manifold X
33. On Abstract Systems
34. What is narrow value learning?
35. A general framework for evaluating aging research. Part 1: reasoning with Longevity Escape Velocity
36. Failures of UDT-AIXI, Part 1: Improper Randomizing
37. Optimizing for Stories (vs Optimizing Reality)
38. How much can value learning be disentangled?
39. [Speech] Worlds That Never Were
40. What is a reasonable outside view for the fate of social movements?
41. Which approach is most promising for aligned AGI?
42. Is Agent Simulates Predictor a "fair" problem?
43. What math do i need for data analysis?
44. What are the components of intellectual honesty?
45. One Website To Rule Them All?
46. Comments on CAIS
47. Why don't people use formal methods?

48. Littlewood's Law and the Global Media
49. "Forecasting Transformative AI: An Expert Survey", Gruetzemacher et al 2019
50. Prediction Contest 2018: Scores and Retrospective

Building up to an Internal Family Systems model

Introduction

Internal Family Systems (IFS) is a psychotherapy school/technique/model which lends itself particularly well for being used alone or with a peer. For years, I had noticed that many of the kinds of people who put in a lot of work into developing their emotional and communication skills, some within the rationalist community and some outside it, kept mentioning IFS.

So I looked at the Wikipedia page about the IFS model, and bounced off, since it sounded like nonsense to me. Then someone brought it up again, and I thought that maybe I should reconsider. So I looked at the WP page again, thought "nah, still nonsense", and continued to ignore it.

This continued until I participated in CFAR mentorship training last September, and we had a class on CFAR's Internal Double Crux (IDC) technique. IDC clicked really well for me, so I started using it a lot and also facilitating it to some friends. However, once we started using it on more emotional issues (as opposed to just things with empirical facts pointing in different directions), we started running into some weird things, which it felt like IDC couldn't quite handle… things which reminded me of how people had been describing IFS. So I finally read up on it, and have been successfully applying it ever since.

In this post, I'll try to describe and motivate IFS in terms which are less likely to give people in this audience the same kind of a "no, that's nonsense" reaction as I initially had.

Epistemic status

This post is intended to give an argument for why *something like* the IFS model *could* be true and a thing that works. It's not really an argument that IFS *is* correct. My reason for thinking in terms of IFS is simply that I was initially super-skeptical of it (more on the reasons of my skepticism later), but then started encountering things which it turned out IFS predicted - and I only found out about IFS predicting those things *after* I familiarized myself with it.

Additionally, I now feel that IFS gives me significantly more gears for understanding the behavior of both other people and myself, and it has been significantly transformative in addressing my own emotional issues. Several other people who I know report it having been similarly powerful for them. On the other hand, aside for a few isolated papers with titles like "proof-of-concept" or "pilot study", there seems to be conspicuously little peer-reviewed evidence in favor of IFS, meaning that we should probably exercise some caution.

I think that, even if not completely correct, IFS is currently the best model that I have for explaining the observations that it's pointing at. I encourage you to read this post

in the style of learning soft skills - trying on this perspective, and seeing if there's anything in the description which feels like it resonates with your experiences.

But before we talk about IFS, let's first talk about building robots. It turns out that if we put together some existing ideas from machine learning and neuroscience, we can end up with a robot design that pretty closely resembles IFS's model of the human mind.

What follows is an intentionally simplified story, which is simpler than *either* the full IFS model *or* a full account that would incorporate everything that I know about human brains. Its intent is to demonstrate that an agent architecture with IFS-style subagents might easily emerge from basic machine learning principles, without claiming that all the details of that toy model would exactly match human brains. A discussion of what exactly IFS *does* claim in the context of human brains follows after the robot story.

Wanted: a robot which avoids catastrophes

Suppose that we're building a robot that we want to be generally intelligent. The hot thing these days seems to be deep reinforcement learning, so we decide to use that. The robot will explore its environment, try out various things, and gradually develop habits and preferences as it accumulates experience. (Just like those human babies.)

Now, there are some problems we need to address. For one, deep reinforcement learning works fine in simulated environments where you're safe to explore for an indefinite duration. However, it runs into problems if the robot is supposed to learn in a real life environment. Some actions which the robot might take will result in catastrophic consequences, such as it being damaged. If the robot is just doing things at random, it might end up damaging itself. Even worse, if the robot does something which could have been catastrophic but narrowly avoids harm, it might then forget about it and end up doing the same thing again!

How could we deal with this? Well, let's look at the existing literature. Lipton et al. (2016) proposed what seems like a promising idea for addressing the part about forgetting. Their approach is to explicitly maintain a memory of *danger states* - situations which are not the catastrophic outcome itself, but from which the learner has previously ended up in a catastrophe. For instance, if "being burned by a hot stove" is a catastrophe, then "being about to poke your finger in the stove" is a danger state. Depending on how cautious we want to be and how many preceding states we want to include in our list of danger states, "going near the stove" and "seeing the stove" can also be danger states, though then we might end up with a seriously stove-phobic robot.

In any case, we maintain a separate storage of danger states, in such a way that the learner never forgets about them. We use this storage of danger states to train a *fear model*: a model which is trying to predict the probability of ending up in a catastrophe from some given novel situation. For example, maybe our robot poked its robot finger at the stove in our kitchen, but poking its robot finger at stoves in other kitchens might be dangerous too. So we want the fear model to generalize from our stove to other stoves. On the other hand, we don't want it to be stove-phobic and run away at the mere sight of a stove. The task of our fear model is to predict exactly how likely it

is for the robot to end up in a catastrophe, given some situation it is in, and then make it increasingly disinclined to end up in the kinds of situations which might lead to a catastrophe.

This sounds nice in theory. On the other hand, Lipton et al. are still assuming that they can train their learner in a simulated environment, and that they can label catastrophic states ahead of time. We don't know in advance every possible catastrophe our robot might end up in - it might walk off a cliff, shoot itself in the foot with a laser gun, be beaten up by activists protesting technological unemployment, or any number of other possibilities.

So let's take inspiration from humans. We can't know beforehand every bad thing that might happen to our robot, but we can identify some classes of things which are correlated with catastrophe. For instance, being beaten or shooting itself in the foot will cause physical damage, so we can install sensors which indicate when the robot has taken physical damage. If these sensors - let's call them "pain" sensors - register a high amount of damage, we consider the situation to have been catastrophic. When they do, we save that situation and the situations preceding it to our list of dangerous situations. Assuming that our robot has managed to make it out of that situation intact and can do anything in the first place, we use that list of dangerous situations to train up a fear model.

At this point, we notice that this is starting to remind us about our experience with humans. For example, the infamous Little Albert experiment. A human baby was allowed to play with a laboratory rat, but each time that he saw the rat, a researcher made a loud scary sound behind his back. Soon Albert started getting scared whenever he saw the rat - and then he got scared of furry things in general.

Something like Albert's behavior could be implemented very simply using something like Hebbian conditioning to get a learning algorithm which picks up on some features of the situation, and then triggers a panic reaction whenever it re-encounters those same features. For instance, it registers that the sight of fur and loud sounds tend to coincide, and then it triggers a fear reaction whenever it sees fur. This would be a basic fear model, and a "danger state" would be "seeing fur".

Wanting to keep things simple, we decide to use this kind of an approach as the fear model of our robot. Also, having read *Consciousness and the Brain*, we remember a few basic principles about how those human brains work, which we decide to copy because we're lazy and don't want to come up with entirely new principles:

- There's a special network of neurons in the brain, called the *global neuronal workspace*. The contents of this workspace are roughly the same as the contents of consciousness.
- We can thus consider consciousness a workspace which many different brain systems have access to. It can hold a single "chunk" of information at a time.
- The brain has multiple different systems doing different things. When a mental object becomes conscious (that is, is projected into the workspace by a subsystem), many systems will synchronize their processing around analyzing and manipulating that mental object.

So here is our design:

- The robot has a hardwired system scanning for signs of catastrophe. This system has several subcomponents. One of them scans the "pain" sensors for signs of

physical damage. Another system watches the “hunger” sensors for signs of low battery.
- Any of these “distress” systems can, alone or in combination, feed a negative reward signal into the global workspace. This tells the rest of the system that this is a bad state, from which the robot should escape.
- If a certain threshold level of “distress” is reached, the current situation is designated as *catastrophic*. All other priorities are suspended and the robot will prioritize getting out of the situation. A memory of the situation and the situations preceding it are saved to a dedicated storage.
- After the experience, the memory of the catastrophic situation is replayed in consciousness for analysis. This replay is used to train up a separate fear model which effectively acts as a new “distress” system.
- As the robot walks around its environment, sensory information about the surroundings will enter its consciousness workspace. When it plans future actions, simulated sensory information about how those actions would unfold enters the workspace. Whenever the new fear model detects features in either kind of sensory information which it associates with the catastrophic events, it will feed “fear”-type “distress” into the consciousness workspace.

So if the robot sees things which remind it of poking at hot stove, it will be inclined to go somewhere else; if it imagines doing something which would cause it to poke at the hot stove, then it will be inclined to imagine doing something else.

Introducing managers

But is this actually enough? We've now basically set up an algorithm which warns the robot when it sees things which have previously preceded a bad outcome. This might be enough for dealing with static tasks, such as not burning yourself at a stove. But it seems insufficient for dealing with things like predators or technological unemployment protesters, who might show up in a wide variety of places and actively try to hunt you down. By the time you see a sign of them, you're already in danger. It would be better if we could learn to avoid them entirely, so that the fear model would never even be triggered.

As we ponder this dilemma, we surf the web and run across this blog post summarizing Saunders, Sastry, Stuhlmüller & Evans (2017). They are also concerned with preventing reinforcement learning agents from running into catastrophes, but have a somewhat different approach. In their approach, a reinforcement learner is allowed to do different kinds of things, which a human overseer then allows or blocks. A separate “blocker” model is trained to predict which actions the human overseer would block. In the future, if the robot would ever take an action which the “blocker” predicts the human overseer would disallow, it will block that action. In effect, the system consists of two separate subagents, one subagent trying to maximize rewards and the other subagent trying to block non-approved actions.

Since our robot has a nice modular architecture into which we can add various subagents which are listening in and taking actions, we decide to take inspiration from this idea. We create a system for spawning dedicated subprograms which try to predict and and block actions which would cause the fear model to be triggered. In theory, this is unnecessary: given enough time, even standard reinforcement learning should learn to avoid the situations which trigger the fear model. But again, trial-and-error can take a very long time to learn exactly which situations trigger fear, so we dedicate a separate subprogram to the task of pre-emptively figuring it out.

Each fear model is paired with a subagent that we'll call a *manager*. While the fear model has associated a bunch of cues with the notion of an impending catastrophe, the manager learns to predict which situations would cause the fear model to trigger. Despite sounding similar, these are not the same thing: one indicates when you are already in danger, the other is trying to figure out what you can do to never end up in danger in the first place. A fear model might learn to recognize signs which technological unemployment protesters commonly wear. Whereas a manager might learn the kinds of environments where the fear model has noticed protesters before: for instance, near the protester HQ.

Then, if a manager predicts that a given action (such as going to the protester HQ) would eventually trigger the fear model, it will block that action and promote some other action. We can use the interaction of these subsystems to try to ensure that the robot only feels fear in situations which already resemble the catastrophic situation so much as to actually *be* dangerous. At the same time, the robot will be unafraid to take safe actions in situations from which it *could* end up in a danger zone, but are themselves safe to be in.

As an added benefit, we can recycle the manager component to also do the same thing as the blocker component in the Saunders et al. paper originally did. That is, if the robot has a human overseer telling it in strict terms not to do some things, it can create a manager subprogram which models that overseer and likewise blocks the robot from doing things which the model predicts that the overseer would disapprove of.

Putting together a toy model

If the robot *does* end up in a situation where the fear model is sounding an alarm, then we want to get it out of the situation as quickly as possible. It may be worth spawning a specialized subroutine just for this purpose. Technological unemployment activists could, among other things, use flamethrowers that set the robot on fire. So let's call these types of subprograms dedicated to escaping from the danger zone, *firefighters*.

So how does the system as a whole work? First, the different subagents act by sending into the consciousness workspace various mental objects, such as an emotion of fear, or an intent to e.g. make breakfast. If several subagents are submitting identical mental objects, we say that they are voting for the same object. On each time-step, one of the submitted objects is chosen at random to become the contents of the workspace, with each object having a chance to be selected that's proportional to its number of votes. If a mental object describing a physical action (an "intention") ends up in the workspace and stays chosen for several time-steps, then that action gets executed by a motor subsystem.

Depending on the situation, some subagents will have more votes than others. E.g. a fear model submitting a fear object gets a number of votes proportional to how strongly it is activated. Besides the specialized subagents we've discussed, there's also a default planning subagent, which is just taking whatever actions (that is, sending to the workspace whatever mental objects) it thinks will produce the greatest reward. This subagent only has a small number of votes.

Finally, there's a self-narrative agent which is constructing a narrative of the robot's actions as if it was a unified agent, for social purposes and for doing reasoning

afterwards. After the motor system has taken an action, the self-narrative agent records this as something like “I, Robby the Robot, made breakfast by cooking eggs and bacon”, transmitting this statement to the workspace and saving it to an episodic memory store for future reference.

Consequences of the model

Is this design any good? Let’s consider a few of its implications.

First, in order for the robot to take physical actions, the intent to do so has to be in its consciousness for a long enough time for the action to be taken. If there are any subagents that wish to prevent this from happening, they must muster enough votes to bring into consciousness some other mental object replacing that intention before it’s been around for enough time-steps to be executed by the motor system. (This is analogous to the concept of the final veto in humans, where consciousness is the last place to block pre-consciously initiated actions before they are taken.)

Second, the different subagents do not see each other directly: they only see the consequences of each other’s actions, as that’s what’s reflected in the contents of the workspace. In particular, the self-narrative agent has no access to information about which subagents were responsible for generating which physical action. It only sees the intentions which preceded the various actions, and the actions themselves. Thus it might easily end up constructing a narrative which creates the internal appearance of a single agent, even though the system is actually composed of multiple subagents.

Third, even if the subagents can’t directly see each other, they might still end up forming alliances. For example, if the robot is standing near the stove, a curiosity-driven subagent might propose poking at the stove (“I want to see if this causes us to burn ourselves again!”), while the default planning system might propose cooking dinner, since that’s what it predicts will please the human owner. Now, a manager trying to prevent a fear model agent from being activated, will eventually learn that if it votes for the default planning system’s intentions to cook dinner (which it saw earlier), then the curiosity-driven agent is less likely to get *its* intentions into consciousness. Thus, no poking at the stove, and the manager’s and the default planning system’s goals end up aligned.

Fourth, this design can make it *really difficult* for the robot to even become aware of the existence of some managers. A manager may learn to support any other mental processes which block the robot from taking specific actions. It does it by voting in favor of mental objects which orient behavior towards anything else. This might manifest as something subtle, such as a mysterious lack of interest towards something that sounds like a good idea in principle, or just repeatedly forgetting to do something, as the robot always seems to get distracted by something else. The self-narrative agent, not having any idea of what’s going on, might just explain this as “Robby the Robot is forgetful sometimes” in its internal narrative.

Fifth, the default planning subagent here is doing something like rational planning, but given its weak voting power, it’s likely to be overruled if other subagents disagree with it (unless some subagents also agree with it). If some actions seem worth doing, but there are managers which are blocking it and the default planning subagent doesn’t have an explicit representation of them, this can manifest as all kinds of procrastinating behaviors and numerous failed attempts for the default planning

system to “try to get itself to do something”, using various strategies. But as long as the managers keep blocking those actions, the system is likely to remain stuck.

Sixth, the purpose of both managers and firefighters is to keep the robot out of a situation that has been previously designated as dangerous. Managers do this by trying to pre-emptively block actions that would cause the fear model agent to activate; firefighters do this by trying to take actions which shut down the fear model agent after it has activated. But the fear model agent activating is not actually the *same* thing as being in a dangerous situation. Thus, both managers and firefighters may fall victim to Goodhart’s law, doing things which block the fear model while being irrelevant for escaping catastrophic situations.

For example, “thinking about the consequences of going to the activist HQ” is something that might activate the fear model agent, so a manager might try to block just *thinking* about it. This has obvious consequence that the robot can’t think clearly about that issue. Similarly, once the fear model has already activated, a firefighter might Goodhart by supporting *any* action which helps activate an agent with a lot of voting power that’s going to think about something entirely different. This could result in compulsive behaviors which were effective at pushing the fear aside, but useless for achieving any of the robot’s actual aims.

At worst, this could cause loops of mutually activating subagents pushing in opposite directions. First, a stove-phobic robot runs away from the stove as it was about to make breakfast. Then a firefighter trying to suppress that fear, causes the robot to get stuck looking at pictures of beautiful naked robots, which is engrossing and thus great for removing the fear of the stove. Then another fear model starts to activate, this one afraid of failure and of spending so much time looking at pictures of beautiful naked robots that the robot won’t accomplish its goal of making breakfast. A separate firefighter associated with this second fear model has learned that focusing the robot’s attention on the pictures of beautiful naked robots *even more* is the most effective action for keeping this new fear temporarily subdued. So the two firefighters are allied and temporarily successful at their goal, but then the first one - seeing that the original stove fear has disappeared - turns off. Without the first firefighter’s votes supporting the second firefighter, the fear manages to overwhelm the second firefighter, causing the robot to rush into making breakfast. This again activates its fear of the stove, but *if* the fear of failure remains strong enough, it might overpower its fear of the stove so that the robot manages to make breakfast in time...

Hmm. Maybe this design isn’t so great after all. Good thing we noticed these failure modes, so that there aren’t any mind architectures like this going around being vulnerable to them!

The Internal Family Systems model

But enough hypothetical robot design; let’s get to the topic of IFS. The IFS model hypothesizes the existence of three kinds of “extreme parts” in the human mind:

- **Exiles** are said to be parts of the mind which hold the memory of past traumatic events, which the person did not have the resources to handle. They are parts of the psyche which have been split off from the rest and are frozen in time of the traumatic event. When something causes them to surface, they tend to flood the mind with pain. For example, someone may have an exile associated with times when they were romantically rejected in the past.

- **Managers** are parts that have been tasked with keeping the exiles permanently exiled from consciousness. They try to arrange a person's life and psyche so that exiles never surface. For example, managers might keep someone from reaching out to potential dates due to a fear of rejection.
- **Firefighters** react when exiles have been triggered, and try to either suppress the exile's pain or distract the mind from it. For example, after someone has been rejected by a date, they might find themselves drinking in an attempt to numb the pain.
- Some presentations of the IFS model simplify things by combining Managers and Firefighters into the broader category of **Protectors**, so only talk about Exiles and Protectors.

Exiles are not limited to being created from the kinds of situations that we would commonly consider seriously traumatic. They can also be created from things like relatively minor childhood upsets, as long as the child didn't feel like they could handle the situation.

IFS further claims that you can treat these parts as something like independent subpersonalities. You can communicate with them, consider their worries, and gradually persuade managers and firefighters to give you access to the exiles that have been kept away from consciousness. When you do this, you can show them that you are no longer in the situation which was catastrophic before, and now have the resources to handle it if something similar was to happen again. This heals the exile, and also lets the managers and firefighters assume better, healthier roles.

As I mentioned in the beginning, when I first heard about IFS, I was turned off by it for several different reasons. For instance, here were some of my thoughts at the time:

1. The whole model about some parts of the mind being in pain, and other parts trying to suppress their suffering. The thing about exiles was framed in terms of *a part of the mind splitting off in order to protect the rest of the mind against damage.* What? That doesn't make any evolutionary sense! A traumatic situation is *just sensory information* for the brain, it's not literal brain damage: it wouldn't have made any sense for minds to evolve in a way that caused parts of it to split off, forcing other parts of the mind to try to keep them suppressed. Why not just… never be damaged in the first place?
2. That whole thing about parts being personalized characters that you could talk to. That… doesn't describe anything in my experience.
3. Also, how does just talking to yourself fix any trauma or deeply ingrained behaviors?
4. IFS talks about everyone having a "True Self". Quote from Wikipedia: *IFS also sees people as being whole, underneath this collection of parts. Everyone has a true self or spiritual center, known as the Self to distinguish it from the parts. Even people whose experience is dominated by parts have access to this Self and its healing qualities of curiosity, connectedness, compassion, and calmness. IFS sees the therapist's job as helping the client to disentangle themselves from their parts and access the Self, which can then connect with each part and heal it, so that the parts can let go of their destructive roles and enter into a harmonious collaboration, led by the Self.* That… again did not sound particularly derived from any sensible psychology.

Hopefully, I've already answered my past self's concerns about the first point. The model itself talks in terms of managers protecting the mind from pain, exiles being exiled from consciousness in order for their pain to remain suppressed, etc. Which is a

reasonable description *of the subjective experience* of what happens. But the evolutionary logic - as far as I can guess - is slightly different: to keep us out of dangerous situations.

The story of the robot describes the actual “design rationale”. Exiles are in fact subagents which are “frozen in the time of a traumatic event”, but they didn’t split off to protect the rest of the mind from damage. *Rather, they were created as an isolated memory block to ensure that the memory of the event wouldn’t be forgotten*. Managers then exist to keep the person away from such catastrophic situations, and firefighters exist to help escape them. Unfortunately, this setup is vulnerable to various failure modes, similar to those that the robot is vulnerable to.

With that said, let’s tackle the remaining problems that I had with IFS.

Personalized characters

IFS suggests that you can experience the exiles, managers and firefighters in your mind as something akin to subpersonalities - entities with their own names, visual appearances, preferences, beliefs, and so on. Furthermore, this isn’t inherently dysfunctional, nor indicative of something like Dissociative Identity Disorder. Rather, even people who are entirely healthy and normal may experience this kind of “multiplicity”.

Now, it’s important to note right off that not everyone has this to a major extent: you don’t *need* to experience multiplicity in order for the IFS process to work. For instance, my parts feel more like bodily sensations and shards of desire than subpersonalities, but IFS still works super-well for me.

In the book *Internal Family Systems Therapy*, Richard Schwartz, the developer of IFS, notes that if a person’s subagents play well together, then that person is likely to feel mostly internally unified. On the other hand, if a person has lots of internal conflict, then they are more likely to experience themselves as having multiple parts with conflicting desires.

I think that this makes a lot of sense, assuming the existence of something like a self-narrative subagent. If you remember, this is the part of the mind which looks at the actions that the mind-system has taken, and then constructs an explanation for why those actions were taken. (See e.g. the posts on the limits of introspection and on the Apologist and the Revolutionary for previous evidence for the existence of such a confabulating subagent with limited access to our true motivations.) As long as all the exiles, managers and firefighters are functioning in a unified fashion, the most parsimonious model that the self-narrative subagent might construct is simply that of a unified self. But if the system keeps being driven into strongly conflicting behaviors, then it can’t necessarily make sense of them from a single-agent perspective. Then it might naturally settle on something like a multiagent approach and experience itself as being split into parts.

Kevin Simler, in Neurons Gone Wild, notes how people with strong addictions seem particularly prone to developing multi-agent narratives:

> *This American Life* did a nice segment on addiction a few years back, in which the producers — seemingly on a lark — asked people to personify their addictions. "It was like people had been waiting all their lives for somebody to ask them this

> question," said the producers, and they gushed forth with descriptions of the 'voice' of their inner addict:
>
> "*The voice is irresistible, always. I'm in the thrall of that voice.*"
>
> "*Totally out of control. It's got this life of its own, and I can't tame it anymore.*"
>
> "*I actually have a name for the voice. I call it Stan. Stan is the guy who tells me to have the extra glass of wine. Stan is the guy who tells me to smoke.*"

This doesn't seem like it explains all of it, though. I've frequently been very dysfunctional, and have always found very intuitive the notion of the mind being split into very parts. Yet I mostly still don't seem to experience my subagents anywhere near as person-like as some others clearly do. I know at least one person who ended up finding IFS because of having all of these talking characters in their head, and who was looking for something that would help them make sense of it. Nothing like that has ever been the case for me: I did experience strongly conflicting desires, but they were just that, strongly conflicting desires.

I can only surmise that it has something to do with the same kinds of differences which cause some people to think mainly verbally, others mainly visually, and others yet in some other hard-to-describe modality. Some fiction writers spontaneously experience their characters as real people who speak to them and will even bother the writer when at the supermarket, and some others don't.

It's been noted that the mechanisms which use to model ourselves and other people overlap - not very surprisingly, since both we and other people are (presumably) humans. So it seems reasonable that some of the mechanisms for representing other people, would sometimes also end up spontaneously recruited for representing internal subagents or coalitions of them.

Why should this technique be useful for psychological healing?

Okay, suppose it's possible to access our subagents somehow. Why would just talking with these entities in your own head, help you fix psychological issues?

Let's consider that a person having exiles, managers and firefighters is costly in the sense of constraining that person's options. If you never want to do anything that would cause you to see a stove, that limits quite a bit of what you can do. I strongly suspect that many forms of procrastination and failure to do things we'd like to do are mostly a manifestation of overactive managers. So it's important not to create those kinds of entities unless the situation really *is* one which should be designated as categorically unacceptable to end up in.

The theory for IFS mentions that not all painful situations turn into trauma: just ones in which we felt helpless and like we didn't have the necessary resources for dealing with it. This makes sense, since if we were capable of dealing with it, then the situation can't have been that catastrophic. The aftermath of the immediate event is important as well: a child who ends up in a painful situation doesn't necessarily end up traumatized, if they have an adult who can put the event in a reassuring context afterwards.

But situations which used to be catastrophic and impossible for us to handle before, aren't necessarily that any more. It seems important to have a mechanism for updating that cache of catastrophic events and for disassembling the protections around it, if the protections turn out to be unnecessary.

How does that process usually happen, without IFS or any other specialized form of therapy?

Often, by talking about your experiences with someone you trust. Or writing about them in private or in a blog.

In my post about *Consciousness and the Brain*, I mentioned that once a mental object becomes conscious, many different brain systems synchronize their processing around it. I suspect that the reason why many people have such a powerful urge to discuss their traumatic experiences with someone else, is that doing so is a way of bringing those memories into consciousness in detail. And once you've dug up your traumatic memories from their cache, their content can be re-processed and re-evaluated. If your brain judges that you now *do* have the resources to handle that event if you ever end up in it again, or if it's something that simply can't happen anymore, then the memory can be removed from the cache and you no longer need to avoid it.

I think it's also significant that, while something like just writing about a traumatic event is sometimes enough to heal, often it's more effective if you have a sympathetic listener who you trust. Traumas often involve some amount of shame: maybe you were called lazy as a kid and are still afraid of others thinking that you are lazy. Here, having friends who accept you and are willing to nonjudgmentally listen while you talk about your issues, is by itself an indication that the thing that you used to be afraid of isn't a danger anymore: there exist people who will stay by your side despite knowing your secret.

Now, when you are talking to a friend about your traumatic memory, you will be going through cached memories that have been stored in an exile subagent. A specific memory circuit - one of several circuits specialized for the act of holding painful memories - is active and outputting its contents into the global workspace, from which they are being turned into words.

Meaning that, in a sense, *your friend is talking directly to your exile.*

Could you hack this process, so that you wouldn't even *need* a friend, and could carry this process out entirely internally?

In my earlier post, I remarked that you could view language as a way of joining two people's brains together. A subagent in your brain outputs something that appears in your consciousness, you communicate it to a friend, it appears in their consciousness, subagents in your friend's brain manipulate the information somehow, and then they send it back to your consciousness.

If you are telling your friend about your trauma, you are in a sense joining your workspaces together, and letting some subagents in *your* workspace, communicate with the "sympathetic listener" subagents in *your friend's* workspace.

So why not let a "sympathetic listener" subagent in your workspace, hook up directly with the traumatized subagents that are also in your own workspace?

I think that something like this happens when you do IFS. You are using a technique designed to activate the relevant subagents in a very specific way, which allows for this kind of a "hooking up" without needing another person.

For instance, suppose that you are talking to a manager subagent which wants to hide the fact that you're bad at something, and starts reacting defensively whenever the topic is brought up. Now, one way by which its activation could manifest, is feeding those defensive thoughts and reactions directly into your workspace. In such a case, you would experience them as your *own* thoughts, and possibly as objectively real. IFS calls this "blending"; I've also previously used the term "cognitive fusion" for what's essentially the same thing.

Instead of remaining blended, you then use various unblending / cognitive defusion techniques that highlight the way by which these thoughts and emotions are coming from a specific part of your mind. You could think of this as wrapping extra content around the thoughts and emotions, and then seeing them through the wrapper (which is obviously not-you), rather than experiencing the thoughts and emotions directly (which you might experience as your own). For example, the IFS book *Self-Therapy* suggests this unblending technique (among others):

> Allow a visual image of the part [subagent] to arise. This will give you the sense of it as a separate entity. This approach is even more effective if the part is clearly a certain distance away from you. The further away it is, the more separation this creates.
>
> Another way to accomplish visual separation is to draw or paint an image of the part. Or you can choose an object from your home that represents the part for you or find an image of it in a magazine or on the Internet. Having a concrete token of the part helps to create separation.

I think of this as something like, you are taking the subagent in question, routing its responses through a visualization subsystem, and then you see a talking fox or whatever. And this is then a representation that your internal subsystems for talking with other people can respond to. You can then have a dialogue with the part (verbally or otherwise) in a way where its responses are clearly labeled as coming from it, rather than being mixed together with all the other thoughts in the workspace. This lets the content coming from the sympathetic-listener subagent and the exile/manager/firefighter subagent be kept clearly apart, allowing you to consider the emotional content as you would as an external listener, preventing you from drowning in it. You're hacking your brain so as to work as the therapist and client as the same time.

The Self

IFS claims that, below all the various parts and subagents, there exists a "true self" which you can learn to access. When you are in this Self, you exhibit the qualities of "calmness, curiosity, clarity, compassion, confidence, creativity, courage, and connectedness". Being at least partially in Self is said to be a prerequisite for working with your parts: if you are not, then you are not able to evaluate their models objectively. The parts will sense this, and as a result, they will not share their models properly, preventing the kind of global re-evaluation of their contents that would update them.

This was the part that I was initially the most skeptical of, and which made me most frequently decide that IFS was not worth looking at. I could easily conceptualize the mind as being made up of various subagents. But then it would just be numerous subagents all the way down, without any single one that could be designated the "true" self.

But let's look at IFS's description of how exactly to get into Self. You check whether you seem to be blended with any part. If you are, you unblend with it. Then you check whether you might also be blended with some other part. If you are, you unblend from it also. You then keep doing this until you can find no part that you might be blended with. All that's left are those "eight Cs", which just seem to be a kind of a global state, with no particular part that they would be coming from.

I now think that "being in Self" represents a state where there no particular subagent is getting a disproportionate share of voting power, and everything is processed by the system as a whole. Remember that in the robot story, catastrophic states were situations in which the organism should *never* end up. A subagent kicking in to prevent that from happening is a kind of a priority override to normal thinking. It blocks you from being open and calm and curious *because* some subagent thinks that doing so would be dangerous. If you then turn off or suspend all those priority overrides, then the mind's default state absent any override seems to be one with the qualities of the Self.

This actually fits at least one model of the function of positive emotions pretty well. Fredrickson (1998) suggests that an important function of positive emotions is to make us engage in activities such as play, exploration, and savoring the company of other people. Doing these things has the effect of building up skills, knowledge, social connections, and other kinds of resources which might be useful for us in the future. If there are no active ongoing threats, then that implies that the situation is pretty safe for the time being, making it reasonable to revert to a positive state of being open to exploration.

The *Internal Family Systems Therapy* book makes a somewhat big deal out of the fact that everyone, even most traumatized people, ultimately has a Self which they can access. It explains this in terms of the mind being organized to protect against damage, and with parts always splitting off from the Self when it would otherwise be damaged. I think the real explanation is much simpler: the mind is not accumulating damage, it is just accumulating a longer and longer list of situations not considered safe.

As an aside, this model feels like it makes me less confused about confidence. It seems like people are really attracted to confident people, and that to some extent it's also possible to fake confidence until it becomes genuine. But if confidence is so attractive and we can fake it, why hasn't evolution just made everyone confident by default?

Turns out that it *has*. The reason why faked confidence gradually turns into genuine confidence is that by forcing yourself to act in confident ways which felt dangerous before, your mind gets information indicating that this behavior is not as dangerous as you originally thought. That gradually turns off those priority overrides that kept you out of Self originally, until you get there naturally.

The reason why being in Self is a requirement for doing IFS, is the existence of conflicts between parts. For instance, recall the stove-phobic robot having a firefighter

subagent that caused it to retreat from the stove into watching pictures of beautiful naked robots. This triggered a subagent which was afraid of the naked-robot-watching preventing the robot from achieving its goals. If the robot now tried to do IFS and talk with the firefighter subagent that caused it to run away from stoves, this might bring to mind content which activated the exile that was afraid of not achieving things. Then *that* exile would keep flooding the mind with negative memories, trying to achieve its priority override of "we need to get out of this situation", and preventing the process from proceeding. Thus, all of the subagents that have strong opinions about the situation need to be unblended from, before integration can proceed.

IFS also has a separate concept of "Self-Leadership". This is a process where various subagents eventually come to trust the Self, so that they allow the person to increasingly remain in Self even in various emergencies. IFS views this as a positive development, not only because it feels nice, but because doing so means that the person will have more cognitive resources available for actually dealing with the emergency in question.

I think that this ties back to the original notion of subagents being generated to invoke priority overrides for situations *which the person originally didn't have the resources to handle*. Many of the subagents IFS talks about seem to emerge from childhood experiences. A child has many fewer cognitive, social, and emotional resources for dealing with bad situations, in which case it makes sense to just categorically avoid them, and invoke special overrides to ensure that this happens. A child's cognitive capacities, models of the world, and abilities to self-regulate are also less developed, so she may have a harder time staying out of dangerous situations *without* having some priority overrides built in. An adult, however, typically has many more resources than a child does. Even when faced with an emergency situation, it can be much better to be able to remain calm and analyze the situation using *all* of one's subagents, rather than having a few of them take over all the decision-making. Thus, it seems to me - both theoretically and practically - that developing Self-Leadership is *really* valuable.

That said, I do not wish to imply that it would be a good goal to *never* have negative emotions. Sometimes blending with a subagent, and experiencing resulting negative emotions, is the right thing to do in that situation. Rather than suppressing negative emotions entirely, Self-Leadership aims to get to a state where any emotional reaction tends to be endorsed by the mind-system as a whole. Thus, if feeling angry or sad or bitter or whatever feels appropriate to the situation, you can let yourself feel so, and then give yourself to that emotion without resisting it. As a result, negative emotions become less unpleasant to experience, since there are fewer subagents trying to fight against them. Also, if it turns out that being in a negative emotional state is no longer useful, the system as a whole can just choose to move back into Self.

Final words

I've now given a brief summary of the IFS model, and explained why I think it makes sense. This is of course not enough to establish the model as *true*. But it might help in making the model plausible enough to at least try out.

I think that most people could benefit from learning and doing IFS on themselves, either alone or together with a friend. I've been saying that exiles/managers/firefighters tend to be generated from trauma, but it's important to realize that these events don't need to be anything immensely traumatic. The kinds of

ordinary, normal childhood upsets that everyone has had can generate these kinds of subagents. Remember, just because you think of a childhood event as trivial *now*, doesn't mean that it felt trivial to you *as a child*. Doing IFS work, I've found exiles related to memories and events which I *thought* left no negative traces, but actually did.

Remember also that it can be really hard to notice the presence of some managers: if they are doing their job effectively, then you might never become aware of them directly. "I don't have any trauma so I wouldn't benefit from doing IFS" isn't necessarily correct. Rather, the cues that I use for detecting a need to do internal work are:

- *Do I have the qualities associated with Self, or is something blocking them?*
- *Do I feel like I'm capable of dealing with this situation rationally, and doing the things which feel like good ideas on an intellectual level?*
- *Do my emotional reactions feel like they are endorsed by my mind-system as a whole, or is there a resistance to them?*

If not, there is often some internal conflict which needs to be addressed - and IFS, combined with some other practices such as Focusing and meditation - has been very useful in learning to solve those internal conflicts.

Even if you don't feel convinced that doing IFS personally would be a good idea, I think adopting its framework of exiles, managers and firefighters is useful for better understanding the behavior of other people. Their dynamics will be easier to recognize in other people if you've had some experience recognizing them in yourself, however.

If you want to learn more about IFS, I would recommend starting with Self-Therapy by Jay Earley. In terms of What/How/Why books, my current suggestions would be:

- How: Self-Therapy by Jay Earley.
- What: Internal Family Systems Therapy, by Richard Schwartz
- Why: The Power of Focusing, by Ann Weiser Cornell (technically not about IFS, but AWC's variant of Focusing gets very close to IFS, and is excellent for conveying the right mindset for it)

This post was written as part of research supported by the Foundational Research Institute. Thank you to everyone who provided feedback on earlier drafts of this article: Eli Tyre, Elizabeth Van Nostrand, Jan Kulveit, Juha Törmänen, Lumi Pakkanen, Maija Haavisto, Marcello Herreshoff, Qiaochu Yuan, and Steve Omohundro.

The 3 Books Technique for Learning a New Skilll

When I'm learning a new skill, there's a technique I often use to quickly gain the basics of the new skill without getting drowned in the plethora of resources that exist. I've found that just 3 resources that cover the skill from 3 separate viewpoints(along with either daily practice or a project) is enough to quickly get all the pieces I need to learn the new skill.

I'm partial to books, so I've called this The 3 Books Technique, but feel free to substitute books for courses, mentors, or videos as needed.

The "What" Book

The "What" book is used as reference material. It should be a thorough resource that gives you a broad overview of your skill. If you run into a novel situation, you should be able to go to this book and get the information you need. It covers the "surface" section of the learning model from nature pictured above.

Positive reviews of this book should contain phrases like "Thorough" and "Got me out of a pinch more than once." Negative reviews of this book should talk about "overwhelming" and "didn't know where to start."

The "How" Book

The "How" Book explains the step-by-step, nuts and bolts of how to put the skill into practice. It often contains processes, tools, and steps. It covers the "deep" part of the learning model covered above.

Positive reviews of this book should talk about "Well structured" and "Clearly thought out." Negative reviews should mention it being "too rote" or "not enough theory."

The "Why" Book

The "WHY" book explains the mindset and intuitions behind the skill. It tries to get into the authors head and lets you understand what to do in novel situations. It should cover the "transfer" part of the learning model above.

Positive reviews of this book should talk about "gaining intuitions" or "really understanding". Negative reviews should contain phrases like "not practical" or "still don't know what steps to take."

The Project or Practice

Once I have these 3 resources, I'll choose a single project or a daily practice that allows me to practice the skills from the "How" book and the mindsets from the "Why" book. If I get stuck, I'll use the "What" book to help me.

Examples

Overcoming Procrastination

"What" Book: The Procrastination Equation by Piers Steel

"How" Book: The Now Habit by Neil Fiore

"Why" Book: The Replacing Guilt blog sequence by Nate Soares

Project or Practice: Five pomodoros every day where I deliberately use the tools from the now habit and the mindsets from replacing guilt. If I find myself stuck, I'll choose from the plethora of techniques in the Procrastination Equation.

Learning Calculus

"What" Book: A First Course in Calculus by Serge Lange

"How" Book: The Khan Academy series on Calculus

"Why" Book: The Essence of Calculus Youtube series by 3blue1brown

Project or Practice: Daily practice of the Khan Academy calculus exercises.

Conclusion

This is a simple technique that I've found very helpful in systematizing my learning process. I would be particularly interested in other skills you've learned and the 3 books you would recommend for those skills.

Some Thoughts on My Psychiatry Practice

I've noticed a marked change in my clientele after going into private practice.[1] Of course I expected class differences-- I charge full fee and don't take insurance. But there are differences that are not as predictable as 'has more money'. During residency I worked at a hospital Medicaid clinic and saw mostly poor, often chronically unemployed people. While monetary problems were a source of stress, they were not nearly as present in people's minds as someone from a middle-class upbringing might think. These people were used to going without. They were not trying to get more. The types of things they talked about were family problems, health problems, and trauma. So much trauma. People's ego-identity crises centered less on their accomplishments and more on their relationships.

The patients I see now are mostly highly successful, highly educated, weathly people, most of whom care a lot about their careers. Their ego-identity crises center around their work and their position in life relative to others. There is a lot of concern about 'the path'. 'Did I go down the right path?' 'Did I make a wrong turn?' There seems to be a great fear of making or having made a wrong decision, which can paralyze their ability to make future decisions. While this group is not without trauma, it is not what they wish to focus on. They will often be dismissive of its effects on them, noting that they clearly got over it in order to get where they are now. Which is, you know, in my office.

Many of my new patients do NOT want to take medication. This is a large change from my patients at the Medicaid clinic who were always requesting more and different pills. And this difference is not because my new patients are less unhappy. They describe intense misery, even a wish to die, going on for months if not years, and yet they struggle through each day in their sisyphean ordeal. They 'power through' until they can't. Until something gives. Then they come to me.

I can think of several good reasons to have concerns about using medication. What are the long-term effects? Could this change my identity? What if this makes me ok with a shitty situation and then I don't fix an underlying problem? But these are not the typical concerns I hear raised. What most of my patients say is that they don't want to 'rely' on a medication. They don't want to be the type of person who takes it. 'That would mean there is something wrong with my brain.' Even though they are clearly very depressed, clearly suffering and hating every day, so long as they can push through without taking a pill they must be 'ok' in some sense. Taking the pill would confirm there is actually something wrong. Taking the pill might mean they are more similar to the patients at the Medicaid clinic than they want to consider.

What struck me about this was how people's desires to assume a certain identity – that of someone who didn't take medication – was more important to them than their actual lived experience. 'This is your life.' And this is broader than to take or not take medication. People will suffer through horrible work situations in order to be the type or person who has that job. 'If your job makes you want to kill yourself, shouldn't you consider quitting it before killing yourself?' 'But I'm good at it.' Identity seems to be everything. Experience is there to tell you if you're on the right way to assuming the

proper identity. If you go through the motions properly you can look the part. What's the difference between looking the part and being the person anyway?

Now refusing medication would be one thing if they wanted to come for weekly therapy and talk through their problems. But many don't. They complain they don't have the time (and it's time, not money that is the concern). They know something is wrong. They were told by their pmd or prior psychiatrist that they should go on an antidepressant. They didn't like the idea, they came to me. For what? I suspect they wanted me to identify the one thing that would help them in one 45 minute session and tell them how to fix it. It doesn't work like that. In this sense, they are not that different from the patients I worked with at the Medicaid clinic. Those patients demanded new meds to fix them, when they clearly had a lot of problems medication was not going to fix. 'This might make you feel less horrible, but it's not going to solve the problems with your marriage.' These new patients eschew being identified in that class, but still in some sense want a 'quick fix'. They want to feel better but keep their illusion of identity intact.

So what's the point of these observations? I'm not quite sure yet. I'm still working that out for myself, which is one of the reasons I decided to write them down. I find I identify more strongly with my current clients, which is unsurprising given we have more similar characteristics and backgrounds. I see some of my own identity struggles in theirs, and it makes me reflect how ridiculous the whole identity struggle is. Everyone is Goodhardting it[2]. All of the time. People want to play a part and they want to be the type of person who plays that part, and their lived experience is a frustrating disappointment which doesn't fit the role they told themselves they have to play. And we live in a society that is vigorously reinforcing 'identity' roles. One where 7 year olds are asked to write essays on their 'identity'. Can we let go of these identity constructs? What is the alternative? Buddhism? Ego death? Self-referential sarcasm? I feel like I'm onto something but not quite there yet. Psychoanalysis is, afterall, an attempt to be more honest with ourselves, and that, it turns out, is much more difficult to do than one might initially think.

[1] * Just noting that I realize that money is not the only factor in the selection process. Patients at the Medicaid clinic were often waiting for months to be seen. A long wait will select against patients that are ambivalent about taking medication. In addition, my website advertises me as being more 'evidence-based', which I think appeals to people who are more likely to have a scientific world-view. Another large difference between my current and former clients is belief in God. Almost none of my current clients believe in God, whereas the majority of my prior clients did. Religion does anticorrelate with class, but I think this is more extreme then one would expect by class alone. I also have a large number of people in finance. How many hedge fund managers are there in NYC anyway? I have many first and second generation immigrants, who have 'pulled myself up by the boot straps' type stories. The wealthy clients I got are 'new money.' Basically I think my advertising captured a demographic that is unusually close to that of my friend/peer group and not necessarily representative of most 'rich people.' The factors that caused them to select me might very well be more relevant than rich v poor in terms of their psychodynamic makeup.

[2] * Goodhardt's law: "When a measure becomes a target, it ceases to be a good measure." In other words – people are optimizing for the superficial qualities by which success is measured, and not the underlying structure of the success.

Book Summary: Consciousness and the Brain

One of the fundamental building blocks of much of consciousness research, is that of Global Workspace Theory (GWT). One elaboration of GWT, which focuses on how it might be implemented in the brain, is the Global Neuronal Workspace (GNW) model in neuroscience. Consciousness and the Brain is a 2014 book that summarizes some of the research and basic ideas behind GNW. It was written by Stanislas Dehaene, a French cognitive neuroscientist with a long background in both consciousness research and other related topics.

The book and its replicability

Given that this is a book on psychology and neuroscience that was written before the replication crisis, an obligatory question before we get to the meat of it is: how reliable are any of the claims in this book? After all, if we think that this is based on research which is probably not going to replicate, then we shouldn't even bother reading the book.

I think that the book's conclusions are at least reasonably reliable in their broad strokes, if not necessarily all the particular details. That is, some of the details in the cited experiments may be off, but I expect most of them to at least be pointing in the right direction. Here are my reasons:

First, scientists in a field usually have an informal hunch of how reliable the different results are. Even before the replication crisis hit, I had heard private comments from friends working in social psychology, who were saying that everything in the field was built on shaky foundations and how they didn't trust even their own findings much. In contrast, when I asked a friend who works with some people doing consciousness research, he reported back that they generally felt that GWT/GNW-style theories have a reasonably firm basis. This isn't terribly conclusive but at least it's a bit of evidence.

Second, for some experiments the book explicitly mentions that they have been replicated. That said, some of the reported experiments seemed to be one-off ones, and I did not yet investigate the details of the claimed replications.

Third, this is a work of cognitive neuroscience. Cognitive neuroscience is generally considered a subfield of cognitive psychology, and cognitive psychology is the part of psychology whose results have so far replicated the best. One recent study tested nine key findings from cognitive psychology, and found that they all replicated. The 2015 "Estimating the reproducibility of Psychological Science" study, managed to replicate 50% of recent results in cognitive psychology, as opposed to 25% of results in social psychology. (If 50% sounds low, remember that we should expect some true results to also fail a single replication, so a 50% replication rate doesn't imply that 50% of the results would be false. Also, a field with a 90% replication rate would probably be too conservative in choosing which experiments to try.) Cognitive psychology replicating pretty well is probably because it deals with phenomena which are much easier to rigorously define and test than social psychology does, so in that regard it's closer to physics than it is to social psychology.

On several occasions, the book reports something like "people did an experiment X, but then someone questioned whether the results of that experiment really supported the hypothesis in question or not, so an experiment X+Y was done that repeated X but also tested Y, to help distinguish between two possible interpretations of X". The general vibe that I get from the book is that different people have different intuitions about how consciousness works, and when someone reports a result that contradicts the intuitions of other researchers, those other researchers are going to propose an alternative interpretation that saves their original intuition. Then people keep doing more experiments until at least one of the intuitions is conclusively disproven - replicating the original experiments in the process.

The analysis of the general reliability of cognitive psychology is somewhat complicated by the fact that these findings are not pure cognitive psychology, but rather cognitive *neuroscience*. Neuroscience is somewhat more removed from just reporting objective findings, since the statistical models used for analyzing the findings can be flawed. I've seen various claims about the problems with statistical tools in neuroscience, but I haven't really dug enough into the field to say to what extent those are a genuine problem.

As suggestive evidence, a lecturer who teaches a "How reliable is cognitive neuroscience?" course reports that before taking a recent iteration of the course, the majority of students

answered the question “If you read about a finding that has been demonstrated across multiple papers in multiple journals by multiple authors, how likely do you think that finding is to be reliable?” as “Extremely likely” and some “Moderately likely”. After taking the course, “Moderately likely” became the most common response with a little under half of the responses, followed by “Slightly likely” by around a quarter of responses and “Extremely likely” with a little over 10% of the responses. Based on this, we might conclude that cognitive neuroscience is moderately reliable, at least as judged by MSc students who’ve just spent time reading and discussing lots of papers critical of cognitive neuroscience.

One thing that’s worth noting is that many of the experiments, including many of the ones this book is reporting on, include two components: a behavioral component and a neuroimaging component. If the statistical models used for interpreting the brain imaging results were flawed, you might get an incorrect impression of what was happening in the brain, but the behavioral results would still be valid. If you’re maximally skeptical of neuroscience, you could choose to throw all of the “inside the brain” results from the book away, and only look at the behavioral results. That seems too conservative to me, but it’s an option. Several of the experiments in the book also use either EEG or single-unit recordings rather than neuroimaging ones; these are much older and simpler techniques than brain imaging is, so are easier to analyze reliably.

So overall, I would expect that the broad strokes of what’s claimed in the book are reasonably correct, even if some of the details might be off.

Defining consciousness

Given that consciousness is a term loaded with many different interpretations, Dehaene reasonably starts out by explaining what he means by consciousness. He distinguishes between three different terms:

- **Vigilance:** whether we “are conscious” in the sense of being awake vs. asleep.
- **Attention:** having focused our mental resources on a specific piece of information.
- **Conscious access:** some of the information we were focusing on, entering our awareness and becoming reportable to others.

For instance, we might be awake (that is, *vigilant*) and staring hard at a computer screen, waiting for some image to be displayed. When that image does get displayed, our *attention* will be on it. But it might flash too quickly for us to report what it looked like, or even for us to realize that it was on the screen in the first place. If so, we don’t have *conscious access* to the thing that we just saw. Whereas if it had been shown for a longer time, we would have conscious access to it.

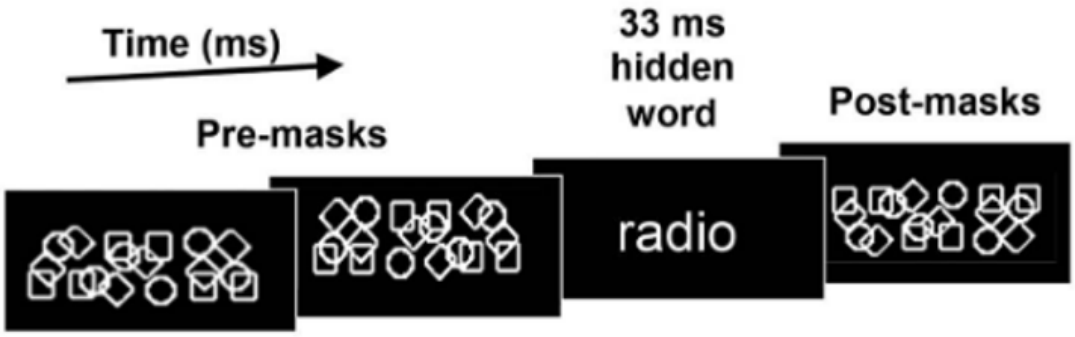
Dehaene says that when he’s talking about consciousness, he’s talking about conscious access, and also that he doesn’t particularly care to debate philosophy and whether this is really *the* consciousness. Rather, since we have a clearly-defined thing which we can investigate using scientific methods, we should just do that, and then think about philosophy once we better understand the empirical side of things.

It seems correct to say that studying conscious access is going to tell us many interesting things, even if it doesn’t solve literally *all* the philosophical questions about consciousness. In the rest of this article, I’ll just follow his conventions and use “consciousness” as a synonym for “conscious access”.

Unconscious processing of meaning

A key type of experiment in Dehaene’s work is *subliminal masking*. Test subjects are told to stare at a screen and report what they see. A computer program shows various geometric

shapes (masks) on the screen. Then at some point, the masks are for a very brief duration replaced with something more meaningful, such as the word “radio”. If the word "radio" is sandwiched between mask shapes, showing it for a sufficiently brief time makes it invisible. The subjects don't even register a brief flicker, as they might if the screen had been totally blank before the word appeared.



By varying the duration for which the word is shown, researchers can control whether or not the subjects see it. Around 40 milliseconds, it is invisible to everyone. Once the duration reaches a certain threshold, which varies somewhat by person but is around 50 milliseconds, the word will be seen around half of the time. When people report not seeing a word, they also fail to name it when asked some time after the trial.

However, even when a masked target doesn’t make it into consciousness, some part of the brain still sees it. It seems as if the visual subsystem started processing the visual stimulus and parsing it in terms of its meaning, but the results of those computations then never made it all the way to consciousness.

One line of evidence for this are subliminal priming experiments, not to be confused with the controversial “social priming” effects in social psychology; unlike those effects, these kinds of priming experiments are well-defined and have been replicated many times. An example of a subliminal priming experiment involves first flashing a hidden word (a prime) so quickly that the participants don’t see it, then following it by a visible word (the target). For instance, people may be primed using the word “radio”, then shown the target word “house”. They are then asked to classify the target word, by e.g. pressing one button if the target word referred to a living thing and another button if it referred to an object.

Subliminal repetition priming refers to the finding that, if the prime and target are the same word and separated by less than a second, then the person will be quicker to classify the target and less likely to make a mistake.

There are indications that when this happens, the brain has parsed some of the prime’s semantic meaning and matched it against the target's meaning. For example, priming works even when the prime is in lower case (radio) and the target is in upper case (RADIO). This might not seem surprising, but look at the difference between e.g. “a” and “A”. These are rather distinct shapes, which we’ve only learned to associate with each other due to cultural convention. Furthermore, while the prime of “range” speeds up the processing of “RANGE”, using “anger” as a prime for “RANGE” has no effect, despite “range” and “anger” having the same letters in a different order. The priming effect comes from the meaning of the prime, rather than just its visual appearance.

The parsing of meaning is not limited to words. If a chess master is shown a simplified chess position for 20 milliseconds, masked so as to make it invisible, they are faster to classify a visible chess position as a check if the hidden position was also a check, and vice versa.

I have reported the above results as saying that the brain does unconscious processing about the meaning of what it sees, but that interpretation has been controversial. After all, something like word processing or identifying a position in check when you have extensive chess experience, is extremely overlearned and could represent an isolated special case rather than showing that the brain processes meaning more generally. The book goes into more detail about the history of this debate and differing interpretations that were proposed; I won't summarize that history in detail, but will just discuss a selection of some experiments which also showed unconscious processing of meaning.

In arithmetic priming experiments, people are first shown a masked single-digit number and then a visible one. They are asked to say whether the target number is larger or smaller than 5. When the number used as a prime is congruent with the target (e.g. smaller than 5 when the target number is also smaller than 5), people respond more quickly than if the two are incongruent. Follow-up work has shown that the effect replicates even if the numbers used as primes are shown in writing ("four") and the target ones as digits ("4"). The priming even works when the prime is an invisible *visual* number and the target a conscious *spoken* number.

Further experiments have shown that the priming effect is the strongest if the prime is the same number as the target number (4 preceding 4). The effect then decreases the more distant the prime is from the target number: 3 preceding 4 shows less of a priming effect, but it still has more of a priming effect than 2 preceding 4 does, and so on. Thus, the brain has done something like extracting an abstract representation of the magnitude of the prime, and used that to influence the processing of the 'target's magnitude.

Numbers could also be argued to be a special case for which we have specialized processing, but later experiments have also shown congruity effects for words in general. For example, when people are shown the word "piano" and asked to indicate whether it is an object or an animal, priming them with a word from a congruent category ("chair") facilitates the correct response, while an incongruent prime ("cat") hinders it.

Some epilepsy patients have had electrodes inserted into their skull for treatment purposes. Some of them have also agreed to have those electrodes used for this kind of research. When they are shown invisible "scary" words such as *danger, rape,* or *poison*, electrodes implanted near the amygdala - the part of the brain involved in fear processing - register an increased level of activation, which is absent for neutral words such as *fridge* or *sonata*.

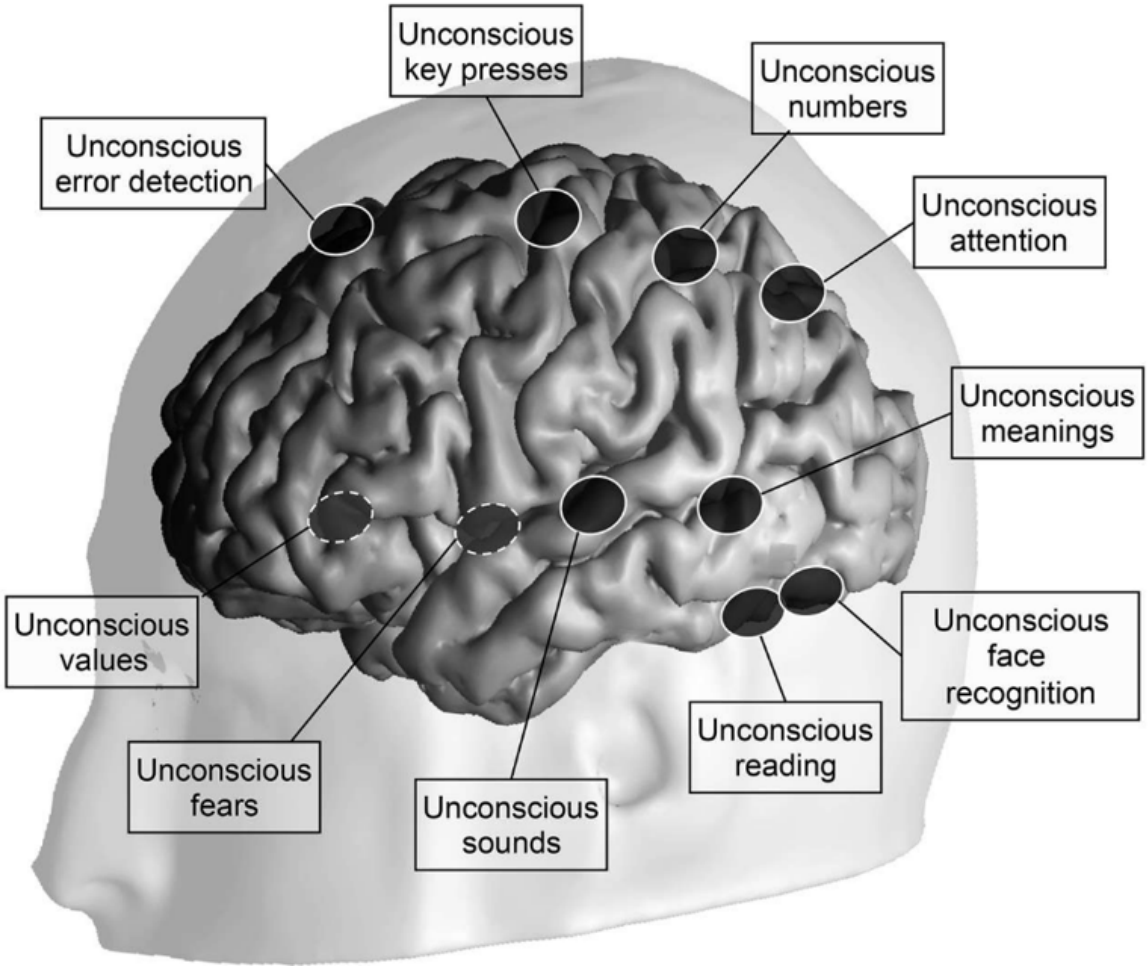
In one study, subjects were shown a "signal", and then had to guess whether to press a button or not press it. As soon as they pressed it, they were told whether they had guessed correctly (earning money) or incorrectly (losing money). Unknown to them, each signal was preceded by a masked shape, which indicated the correct response: one kind of a shape indicated that pressing the button would earn them money, another shape indicated that not pressing the button would earn them money, and a third one meant that either choice had an equal chance of being correct. Even though the subjects were never aware of seeing the shape, once enough trials had passed, they started getting many more results correct than chance alone would indicate. An unconscious value system had associated the shapes with different actions, and was using the subliminal primes for choosing the right action.

Unconscious processing can also weigh the average value of a number of variables. In one type of experiment, subjects are choosing cards from four different decks. Each deck has cards that cause the subject to either earn or lose reward money, with each deck having a different distribution of cards. Two of the decks are "bad", causing the subjects to lose money on net, and two of them are "good", causing them to gain money on net. By the end of the experiment, subjects have consciously figured out which one is which, and can easily report this. However, measurements of skin conductance indicate that even before they have consciously figured out the good and bad decks, there comes a point when they've pulled enough cards that being about to draw a card from a bad deck causes their hands to

sweat. A subconscious process has already started generating a prediction of which decks are bad, and is producing a subliminal gut feeling.

A similar unconscious averaging of several variables can also be shown using the subliminal priming paradigm. Subjects are shown five arrows one at a time, some of which point left and some of which point right. They are then asked for the direction that the majority of the arrows were pointing to. When the arrows are made invisible by subliminal masking, subjects who are forced to guess feel like they are just making random guesses, but are in fact responding much more accurately than by chance alone.

There more examples in the book, but these should be enough to convey the general idea: that many different sensory inputs are automatically registered and processed in the brain, even if they are never shown for long enough to make it all the way to consciousness. Unconsciously processed stimuli can even cause movement commands to be generated in the motor cortex and sent to the muscles, though not necessarily at an intensity which would sufficient to cause actual action.



What about consciousness, then?

So given everything that our brain does automatically and without conscious awareness, what’s up with consciousness? What is it, and what does it do?

Some clues can be found from investigating the neural difference between conscious and unconscious stimuli. Remember that masking experiments show a threshold effect in whether a stimulus is seen or not: if a stimulus which is preceded by a mask is shown for 40 milliseconds, then it's invisible, but around 50 milliseconds it starts to become visible. In experiments where the duration of the stimulus is carefully varied, there is an all-or-nothing effect: subjects do not report seeing more and more of the stimulus as the duration is gradually increased. Rather they either see it in its entirety, or they see nothing at all.

A key finding, replicated across different sensory modalities and different methods for measuring brain activation (fMRI, EEG, and MEG) is that a stimulus becoming conscious involves an effect where, once the strength of a stimulus exceeds a certain threshold, the neural signal associated with that stimulus is massively boosted and spreads to regions in the brain which it wouldn't have reached otherwise. Exceeding the key threshold causes the neural signal generated by the sensory regions to be amplified, with the result that the associated signal could be spread more widely, rather than fading away before it ever reached all the regions.

Dehaene writes, when discussing an experiment where this was measured using visually flashed words as the stimulus:

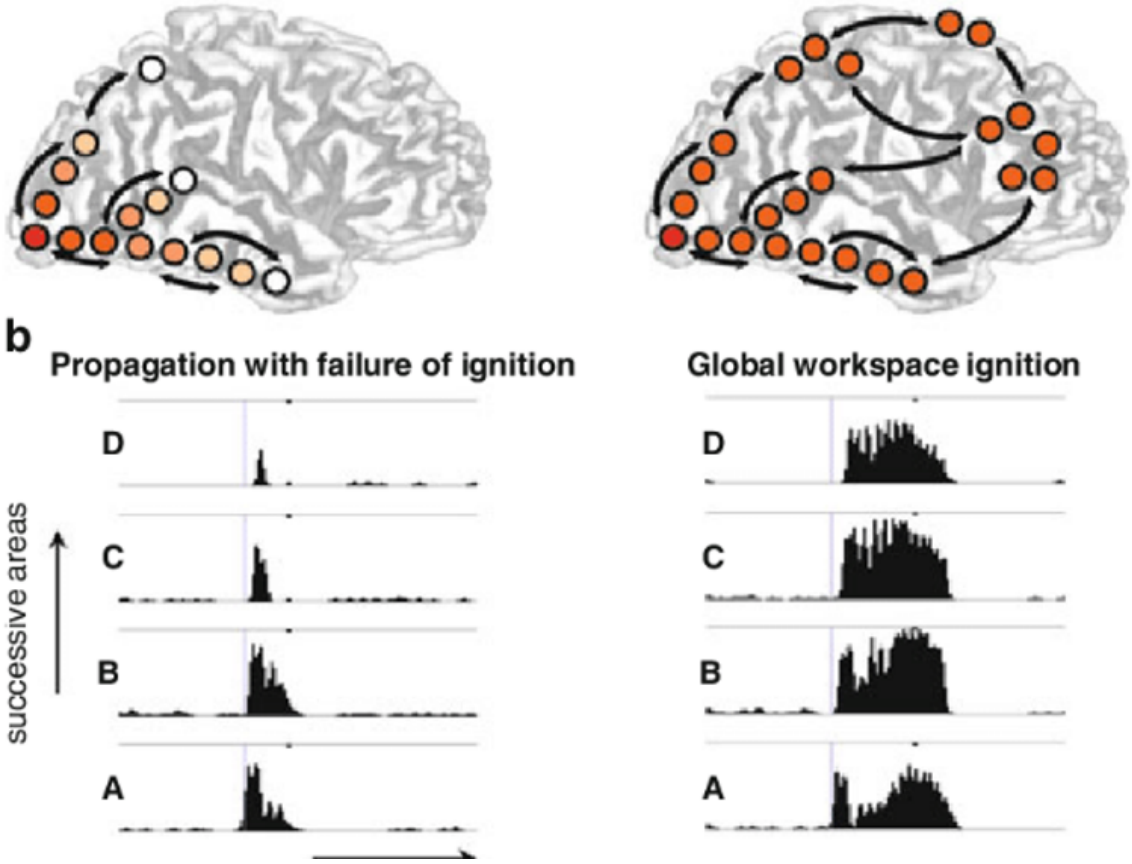
> By measuring the amplitude of this activity, we discovered that the amplification factor, which distinguishes conscious from unconscious processing, varies across the successive regions of the visual input pathway. At the first cortical stage, the primary visual cortex, the activation evoked by an unseen flashed word is strong enough to be easily detectable. However, as it progresses forward into the cortex, masking makes it lose strength. Subliminal perception can thus be compared to a surf wave that looms large on the horizon but merely licks your feet when it reaches the shore. By comparison, conscious perception is a tsunami—or perhaps an avalanche is a better metaphor, because conscious activation seems to pick up strength as it progresses, much as a minuscule snowball gathers snow and ultimately triggers a landslide.
>
> To bring this point home, in my experiments I flashed words for only 43 milliseconds, thereby injecting minimal evidence into the retina. Nevertheless, activation progressed forward and, on conscious trials, ceaselessly amplified itself until it caused a major activation in many regions. Distant brain regions also became tightly correlated: the incoming wave peaked and receded simultaneously in all areas, suggesting that they exchanged messages that reinforced one another until they turned into an unstoppable avalanche. Synchrony was much stronger for conscious than for unconscious targets, suggesting that correlated activity is an important factor in conscious perception.
>
> These simple experiments thus yielded a first signature of consciousness: an amplification of sensory brain activity, progressively gathering strength and invading multiple regions of the parietal and prefrontal lobes. This signature pattern has often been replicated, even in modalities outside vision. For instance, imagine that you are sitting in a noisy fMRI machine. From time to time, through earphones, you hear a brief pulse of additional sound. Unknown to you, the sound level of these pulses is carefully set so that you detect only half of them. This is an ideal way to compare conscious and unconscious perception, this time in the auditory modality. And the result is equally clear: unconscious sounds activate only the cortex surrounding the primary auditory area, and again, on conscious trials, an avalanche of brain activity amplifies this early sensory activation and breaks into the inferior parietal and prefrontal areas

Dehaene goes into a considerable amount of detail about the different neuronal signatures which have been found to correlate with consciousness, and the experimental paradigms which have been used to test whether or not those signatures are mere correlates rather than parts of the causal mechanism. I won't review all of that discussion here, but will summarize some of his conclusions.

Consciousness involves a neural signal activating self-reinforcing loops of activity, which causes wide brain regions to synchronize to process that signal.

Consider what happens when someone in the audience of a performance starts clapping their hands, soon causing the whole audience to burst into applause. As one person starts clapping, other people hear it and start clapping in turn; this becomes a self-reinforcing effect where your clapping causes other people to clap, and you are more likely to continue clapping if other people are also still clapping. In a similar way, the threshold effect of conscious activation seems to involve some neurons sending a signal, causing other neurons to activate and join in on broadcasting that signal. The activation threshold is a point where enough neurons have sufficient mutual excitation to create a self-sustaining avalanche of excitation, spreading throughout the brain.

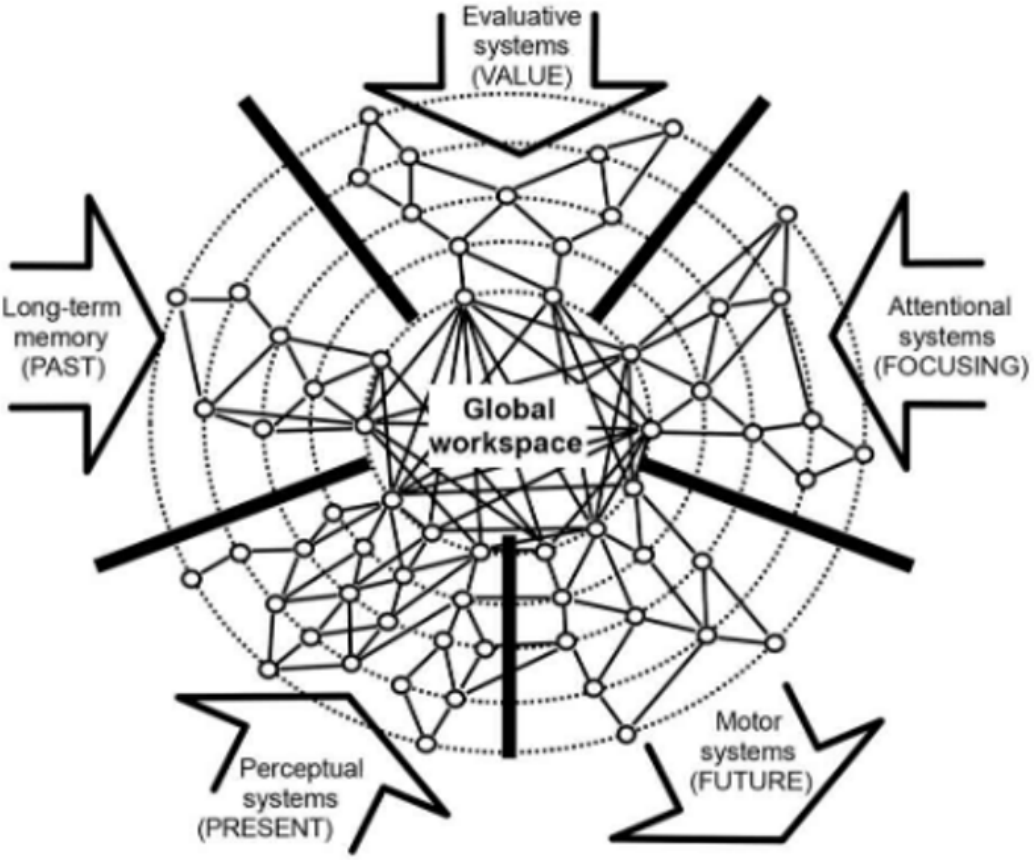
The spread of activation is further facilitated by a “brain web” of long-distance neurons, which link multiple areas of the brain together into a densely interconnected network. Not all areas of the brain are organized in this way: for instance, sensory regions are mostly connected to their immediate neighbors, with visual area V1 being primarily only connected to visual area V2, and V2 mostly to V1 and V3, and so on. But higher areas of the cortex are much more joined together, in a network where area A projecting activity to area B usually means that area B also projects activity back to area A. They also involve triangular connections, where region A might project into regions B and C, which then both also project to each other and back to A. This long-distance network joins not only areas of the cortex, but is also connected to regions such as the thalamus (associated with e.g. attention and vigilance), the basal ganglia (involved in decision-making and action), and the hippocampus (involved in episodic memory).



A stimulus becoming conscious involves the signal associated with it achieving enough strength to activate some of the associative areas that are connected with this network, which Dehaene calls the “global neuronal workspace” (GNW). As those areas start broadcasting the signal associated with the stimulus, other areas in the network receive it and start broadcasting it in turn, creating the self-sustaining loop. As this happens, many different regions will end up processing the signal at the same time, synchronizing their processing around the contents of that signal. Dehaene suggests that the things that we are conscious of at any given moment, are exactly the things which are being processed in our GNW at that moment.

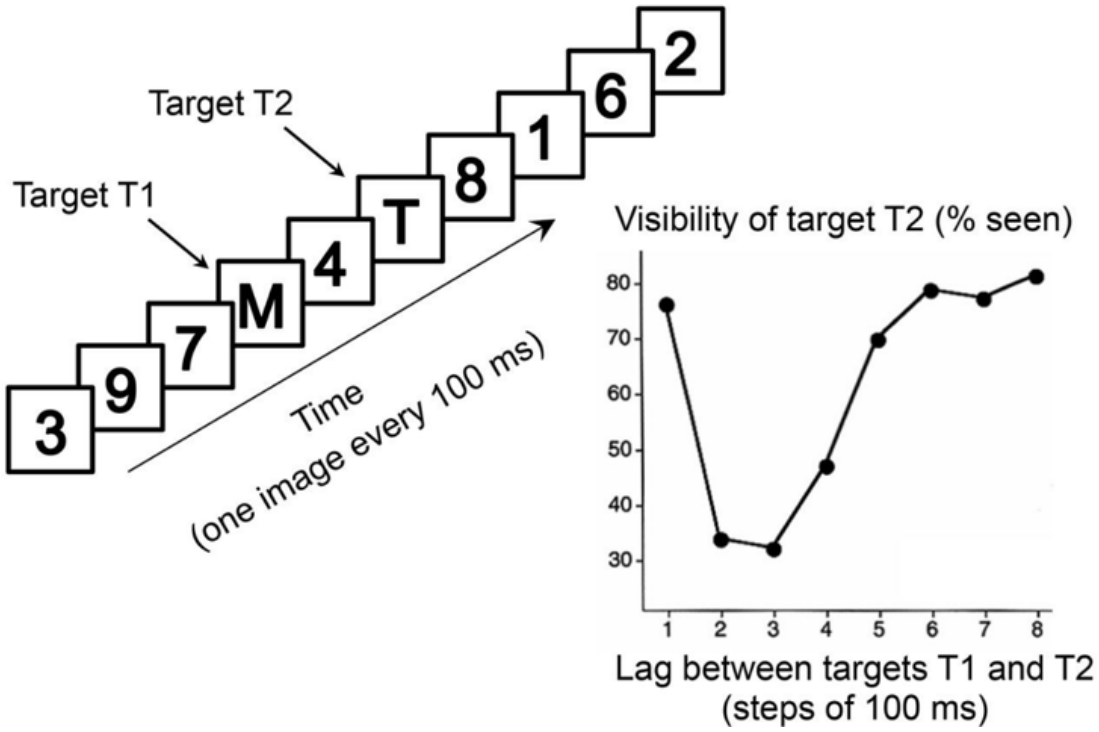
Dehaene describes this as “a decentralized organization without a single physical meeting site” where “an elitist board of executives, distributed in distant territories, stays in sync by exchanging a plethora of messages”. While he mostly reviews evidence gathered from investigating sensory inputs, his model holds that besides sensory areas, many other regions - such as the ones associated with memory and attention - also feed into and manipulate the contents of the network. Once a stimulus enters the GNW, networks regulating top-down attention can amplify and “help keep alive” stimuli which seems especially important to focus on, and memory networks can commit the stimulus into memory, insert into the network earlier memories which were triggered by the sight of the stimulus, or both.

In the experiments on subliminal processing, an unconscious prime may affect the processing of a conscious stimulus that comes very soon afterwards, but since its activation soon fades out, it can’t be committed to memory or verbally reported on afterwards. A stimulus becoming conscious and being maintained in the GNW, both keeps its signal alive for longer, and also allows it better access to memory networks which may store it in order for it to be re-broadcast into the GNW later.

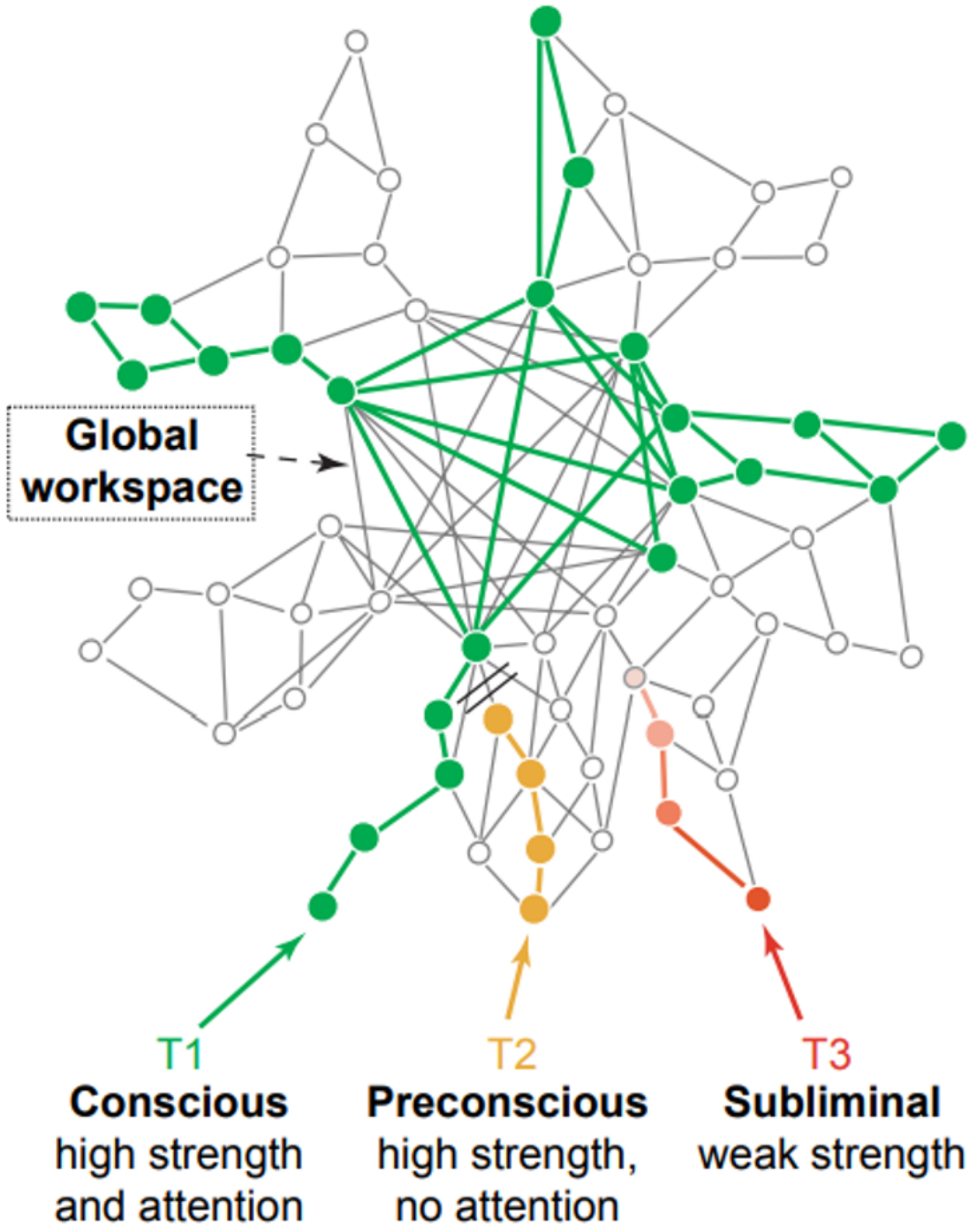


The global workspace can only be processing a single item at a time.

Various experiments show the existence of an “attentional blink”: if your attention is strongly focused on one thing, it takes some time to disengage from it and reorient your attention to something else. For instance, in one experiment people are shown a stream of symbols. Most of the symbols are digits, but some are letters. People are told to remember the letters. While the first letter is easy to remember, if two letters are shown in rapid succession, the subjects might not even realize that two of them were present - and they might be surprised to learn that this was the case. The act of attending to the first letter enough to memorize it, creates a “blink of the mind” which prevents the second letter from ever being noticed.



Dehaene’s explanation for this is that the GNW can only be processing a single item at once. The first letter is seen, processed by the early visual centers, then reaches sufficient strength to make it into the workspace. This causes the workspace neurons to synchronize their processing around the first letter and try to keep the signal active for long enough for it to be memorized - and while they are still doing so, the second letter shows up. It is also processed by the visual regions and makes it to the associative region, but the attention networks are still reinforcing the signal associated with the original letter and keeping it active in the workspace. The new letter can’t muster enough activation in time to get its signal broadcast into the workspace, so by the time the activation generated by the first letter starts to fade, the signal from the second letter has also faded out. As a result, the second signal never makes it to the workspace where it could leave a conscious memory trace of having been observed.



When two simultaneous events happen, it doesn't *always* mean that awareness of the other one is suppressed. If there isn't too much distraction - due to "internal noise, distracting thoughts, or other incoming stimuli" - the signal of the second event may survive for long enough in an unconscious buffer, making it to the GNW after the first event has been processed. The use of a post-stimulus masking shape in the subliminal masking experiments helps erase the contents of this buffer, by providing a new stimulus that overwrites the old one. In these cases, people's judgment of the timing of the events is systematically wrong: rather than experiencing the events to have happened simultaneously, they believe the second event to have happened at the time when the event entered their consciousness.

As an interesting aside, as a result of these effects, the content of our consciousness is always slightly delayed relative to when an event actually happened - a stimulus getting into the GNW takes at least one-third of a second, and may take substantially longer if we are distracted. The brain contains a number of mechanisms for compensating the delay in GNW access, such as prediction mechanisms which anticipate how familiar events should happen before they've actually happened.

Disrupting or stimulating the GNW, has the effects that this theory would predict.

One of the lines of argument by which Dehaene defends the claim that GNW activity is genuinely the same thing as conscious activity, and not a more correlate, is that artificially interfering with GNW activity has the kinds of effects that we might expect.

To do this, we can use Transcranial Magentic Stimulation (TMS) to create magnetic fields which stimulate electric activity in the brain, or if electrodes have been placed in a person's brain, those can be used to stimulate the brain directly.

In one experiment, TMS was used to stimulate the visual cortex of test subjects, in a way that created a hallucination of light. By varying the intensity of the stimulation, the researchers could control whether or not the subjects noticed anything. On trials when the subjects reported becoming conscious of a hallucination, an avalanche wave associated with consciousness popped up, reaching consciousness faster than normal. In Dehaene's interpretation, the magnetic pulse bypassed the normal initial processing stages for vision and instead created a neuronal activation directly at a higher cortical area, speeding up conscious access by about 0.1 seconds.

Experiments have also used TMS to successfully erase awareness of a stimulus. One experiment described in the book uses a dual TMS setup. First, a subject is zapped with a magnetic pulse that causes them to see a bit of (non-existent) movement. After it has been confirmed that subjects report becoming conscious of movement when they are zapped with the first pulse, they are then subjected to a trial where they are first zapped with the same pulse, then immediately thereafter with another pulse that's aimed to disrupt the signal from getting access to the GNW. When this is done, subjects report no longer being aware of having seen any movement.

The functions of consciousness

So what exactly *is* the function of consciousness? Dehaene offers four different functions.

Conscious sampling of unconscious statistics and integration of complicated information

Suppose that you a Bayesian decision theorist trying to choose between two options, A and B. For each two options, you've computed a probability distribution about the possible outcomes that may result if you choose either A or B. In order to actually make your choice, you need to collapse your probability distributions into a point estimate of the expected value of choosing A versus B, to know which one is actually better.

In Dehaene's account, consciousness does something like this. We have a number of unconscious systems which are constantly doing Bayesian statistics and constructing probability distributions about how to e.g. interpret visually ambiguous stimuli, weighing multiple hypotheses at the same time. In order for decision-making to actually be carried

out, the system has to choose one of the interpretations, and act based on the assumption of that interpretation being correct. The hypothesis that the unconscious process selects as correct, is then what gets fed into consciousness. For example, when I look at the cup of tea in front of me, I don't see a vast jumble of shifting hypotheses of what this visual information might represent: rather, I just see what I think is a cup of tea, which is what a subconscious process has chosen as the most likely interpretation.

Dehaene offers the analogy of the US President being briefed by the FBI. The FBI is a vast organization, with thousands of employees: they are constantly shifting through enormous amounts of data, and forming hypotheses about topics which have national security relevance. But it would be useless for the FBI to present to the President every single report collected by every single field agent, as well as every analysis compiled by every single analyst in response. Rather, the FBI needs to internally settle on some overall summary of what they believe is going on, and then present that to the President, who can then act based on the information. Similarly, Dehaene suggests that consciousness is a place where different brain systems can exchange summaries of their models, and to integrate conflicting evidence in order to arrive to an overall conclusion.

Dehaene discusses a few experiments which lend support this interpretation, though here the discussion seems somewhat more speculative than in other parts of the book. One of his pieces of evidence is of recordings of neuronal circuits which integrate many parts of a visual scene into an overall image, resolving local ambiguities by using information from other parts of the image. Under anesthesia, neuronal recordings show that this integration process is disrupted; consciousness "is needed for neurons to exchange signals in both bottom-up and top-down directions until they agree with each other". Another experiment shows that if people are shown an artificial stimulus which has been deliberately crafted to be ambiguous, people's conscious impression of the correct interpretation keeps shifting: first it's one interpretation, then the other. By varying the parameters of the stimulus, researchers can control roughly how often people see each interpretation. If Bayesian statistics would suggest that interpretation A was 30% likely and interpretation B 70% likely, say, then people's impression of the image will keep shifting so that they will see interpretation A roughly 30% of the time and interpretation B roughly 70% of the time.

> What we see, at any time, tends to be the most likely interpretation, but other possibilities occasionally pop up and stay in our conscious vision for a time duration that is proportional to their statistical likelihood. Our unconscious perception works out the probabilities - and then our consciousness samples from them at random.

In Dehaene's account, consciousness is involved in higher-level integration of the meaning of concepts. For instance, our understanding of a painting such as the Mona Lisa is composed of many different things. Personally, if I think about the Mona Lisa, I see a mental image of the painting itself, I get an association with the country of Italy, I remember having first learned about the painting in a Donald Duck story, and I also remember my friend telling me about the time she saw the original painting itself. These are different pieces of information, stored in different formats in different regions of the brain, and the kind of global neuronal integration carried out by the GNW allows all of these different interpretations to come together, with every system participating in constructing an overall coherent, synchronous interpretation.

All of this sounds sensible enough. At the same, after all the previous discussion about unconscious decision-making and unconscious integration of information, this leaves me feeling somewhat unsatisfied. If it has been shown that e.g. unconsciously processed cues are enough to guide our decision-making, then how do we square that with the claim that consciousness is necessary for settling on a single interpretation that would allow us to take actions?

My interpretation is that even though unconscious processing and decision-making happens, its effect is relatively weak. If you prime people with a masked stimulus, then that influences

their decision-making so as to give them better performance - but it doesn't give them *perfect* performance. In the experiment where masked cues predicted the right action and unconscious learning associated each cue with the relevant action, the subjects only ended up with an average of 63% correct actions.

Looking at the cited paper itself, the authors themselves note that if the cues had been visible, it would only have taken a couple of trials for the subjects to learn the optimal behaviors. In the actual experiment, their performance slowly improved until it reached a plateau around 60 trials. Thus, even though unconscious learning and decision-making happens, conscious learning and decision-making can be significantly more effective.

Second, while I don't see Dehaene mentioning it, I've always liked the PRISM theory of consciousness, which suggests that one of the functions of consciousness is to be a place for resolving conflicting plans for controlling the skeletal muscles. In the unconscious decision-making experiments, the tasks have mostly been pretty simple, and only involved the kinds of goals that could all be encapsulated within a single motivational system. In real life however, we often run into situations where different brain systems output conflicting instructions. For instance, if we are carrying a hot cup of tea, our desire to drop the cup may be competing against our desire to carry it to the table, and these may have their origin in very different sorts of motivations. Information from both systems would need to be taken into account and integrated in order to make an overall decision.

To stretch Dehaene's FBI metaphor: as long as the FBI is doing things that fall within their jurisdiction and they are equipped to handle, then they can just do that without getting in contact with the President. But if the head of the FBI and the head of the CIA have conflicting ideas about what should be done, on a topic on which the two agencies have overlapping jurisdiction, then it might be necessary to bring the disagreement out in the open so that a higher-up can make the call. Of course, there isn't any single "President" in the brain who would make the final decision: rather, it's more like the chiefs of all the other alphabet soup bureaus were also called in, and they then hashed out the details of their understanding until they came to a shared agreement about what to do.

Lasting thoughts and working memory

As already touched upon, consciousness is associated with memory. Unconsciously registered information tends to fade very quickly and then disappear. In all the masking experiments, the duration between the prime and the target is very brief; if the duration would be any longer, there would be no learning or effect on decision-making. For e.g. associating cues and outcomes with each other over an extended period of time, the cue has to be consciously perceived.

Dehaene describes an experiment which demonstrates exactly this:

> The cognitive scientists Robert Clark and Larry Squire conducted a wonderfully simple test of temporal synthesis: time-lapse conditioning of the eyelid reflex. At a precisely timed moment, a pneumatic machine puffs air toward the eye. The reaction is instantaneous: in rabbits and humans alike, the protective membrane of the eyelid immediately closes. Now precede the delivery of air with a brief warning tone. The outcome is called Pavlovian conditioning (in memory of the Russian physiologist Ivan Petrovich Pavlov, who first conditioned dogs to salivate at the sound of a bell, in anticipation of food). After a short training, the eye blinks to the sound itself, in anticipation of the air puff. After a while, an occasional presentation of the isolated tone suffices to induce the "eyes wide shut" response.
>
> The eye-closure reflex is fast, but is it conscious or unconscious? The answer, surprisingly, depends on the presence of a temporal gap. In one version of the test, usually termed "delayed conditioning," the tone lasts until the puff arrives. Thus the two

stimuli briefly coincide in the animal's brain, making the learning a simple matter of coincidence detection. In the other, called "trace conditioning," the tone is brief, separated from the subsequent air puff by an empty gap. This version, although minimally different, is clearly more challenging. The organism must keep an active memory trace of the past tone in order to discover its systematic relation to the subsequent air puff. To avoid any confusion, I will call the first version "coincidence-based conditioning" (the first stimulus lasts long enough to coincide with the second, thus removing any need for memory) and the second "memory-trace conditioning" (the subject must keep in mind a memory trace of the sound in order to bridge the temporal gap between it and the obnoxious air puff).

The experimental results are clear: coincidence-based conditioning occurs unconsciously, while for memory-trace conditioning, a conscious mind is required. In fact, coincidence-based conditioning does not require any cortex at all. A decerebrate rabbit, without any cerebral cortex, basal ganglia, limbic system, thalamus, and hypothalamus, still shows eyelid conditioning when the sound and the puff overlap in time. In memory-trace conditioning, however, no learning occurs unless the hippocampus and its connected structures (which include the prefrontal cortex) are intact. In human subjects, memory-trace learning seems to occur if and only if the person reports being aware of the systematic predictive link between the tone and the air puff. Elderly people, amnesiacs, and people who were simply too distracted to notice the temporal relationship show no conditioning at all (whereas these manipulations have no effect whatsoever on coincidence-based conditioning). Brain imaging shows that the subjects who gain awareness are precisely those who activate their prefrontal cortex and hippocampus during the learning.

Carrying out artificial serial operations

Consider what happens when you calculate 12 * 13 in your head.

When you do so, you have some conscious awareness of the steps involved: maybe you first remember that 12 * 12 = 144 and then add 144 + 12, or maybe you first multiply 12 * 10 = 120 and then keep that result in memory as you multiply 12 * 3 = 36 and then add 120 + 36. Regardless of the strategy, the calculation happens consciously.

Dehaene holds that this kind of multi-step arithmetic can't happen unconsciously. We can do *single-step* arithmetic unconsciously: for example, people can be shown a single masked digit *n*, and then be asked to carry out one of three operations. People might be asked to name the digit (the "*n*" task), to add 2 to *n* and report the resulting the number (the "*n* + 2" task), or to report whether or not it's smaller than 5 (the "*n* < 5" task). On all of these tasks, even if people haven't consciously seen the digit, when they are forced to guess they typically get the right answer half of the time.

However, unconscious *two-step* arithmetic fails. If people are flashed an invisible digit and told to first add 2 to it, and then report whether the result is more or less than 5 (the "(*n* + 2) > 5" task), their performance is on the chance level. The unconscious mind can carry out a single arithmetic operation, but it can't then store the result of that operation and use it as the input of a second operation, even though it could carry out either of the two operations alone.

Dehaene notes that this might seem to contradict a previous finding, which is that the unconscious brain can *accumulate* multiple pieces of information over time. For instance, in the arrow experiment, people were shown several masked arrows one at a time; at the end, they could tell whether most of them had been pointing to the left or to the right. Dehaene says that the difference is that opening a neural circuit which accumulates multiple observations is a single operation for the brain: and while the accumulator stores information

of how many arrows have been observed so far, that information can't be taken out of it and used as an input for a second calculation.

The accumulator also can't reach a decision by itself: for instance, if people saw the arrows consciously, they could reach a decision after having seen three arrows that pointed one way, knowing that the remaining arrows couldn't change the overall decision anymore. In unconscious trials, they can't use this kind of strategic reasoning: the unconscious circuit can only keep adding up the arrows, rather than adding up the arrows *and* also checking whether a rule of type "if seen_arrows > 3" has been satisfied yet.

According to Dehaene, implementing such rules is one of the functions of consciousness. In fact, he explicitly compares consciousness to a production system: an AI design which holds a number of objects in a working memory, and also contains a number of IF-THEN rules, such as "if there is an A in working memory, change it to sequence BC". If multiple rules match, one of them is chosen for execution according to some criteria. After one of the rules has fired, the contents of the working memory gets updated, and the cycle repeats. The conscious mind, Dehaene says, works using a similar principle - creating a biological Turing machine that can combine operations from a number of neuronal modules, flexibly chaining them together for serial execution.

A social sharing device

If a thought is conscious, we can describe it and report it to other people using language. I won't elaborate on this, given that the advantages of being able to use language to communicate with others are presumably obvious. I'll just note that Dehaene highlights one interesting perspective: one where other people are viewed as additional modules that can carry out transformations on the objects in the workspace.

Whether it's a subsystem in the brain that's applying production rules to the workspace contents, or whether you are communicating the contents to another person who then comments on it (as guided by some subsystem in *their* brain), the same principle of "production rules transforming the workspace contents" still applies. Only in one of the cases, the rules and transformations come from subsystems that are located within a single brain, and in the other case subsystems from multiple brains are engaged in joint manipulation of the contents - though of course the linguistic transmission is lossy, since subsystems in multiple brains can't communicate with the same bandwidth as subsystems in a single brain. (Yet.)

Other stuff

Dehaene also discusses a bunch of other things in his book: for instance, he talks about comatose patients and how his research has been applied to study their brains, in order to predict which patients will eventually recover and which ones will remain permanently unresponsive. This is pretty cool, and feels like a confirmation of the theories being on the right track, but since it's no longer elaborating on the mechanisms and functions of consciousness, I won't cover that here.

Takeaways for the rest of the sequence

This has been a pretty long post. Now that we're at the end, I'm just going to highlight a few of the points which will be most important when we go forward in the multiagent minds sequence:

- Consciousness can only contain a single mental object at a time.

- The brain has multiple different systems doing different things; many of the systems do unconscious processing of information. When a mental object becomes conscious, many systems will synchronize their processing around analyzing and manipulating that mental object.
- The brain can be compared to a production system, with a large number of specialized rules which fire in response to specific kinds of mental objects. E.g. when doing mental arithmetic, applying the right sequence of arithmetic operations for achieving the main goal.

If we take the view of looking at various neural systems as being literally technically subagents, then we can reframe the above points as follows:

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

Next up: constructing a mechanistic sketch of how a mind might work, combining the above points as well as the kinds of mechanisms that have already been demonstrated in contemporary machine learning, to finally end up with a model that pretty closely resembles the Internal Family Systems one.

Disentangling arguments for the importance of AI safety

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

Note: my views have shifted significantly since writing this post. I now consider items 1, 2, 3, and 6.2 to be different facets of one core argument, which I call the "second species" argument, and which I explore in depth in this report. And I don't really think of 4 as an AI safety problem any more.

I recently attended the 2019 Beneficial AGI conference organised by the Future of Life Institute. I'll publish a more complete write-up later, but I was particularly struck by how varied attendees' reasons for considering AI safety important were. Before this, I'd observed a few different lines of thought, but interpreted them as different facets of the same idea. Now, though, I've identified at least 6 distinct serious arguments for why AI safety is a priority. By distinct I mean that you can believe any one of them without believing any of the others - although of course the particular categorisation I use is rather subjective, and there's a significant amount of overlap. In this post I give a brief overview of my own interpretation of each argument (note that I don't necessarily endorse them myself). They are listed roughly from most specific and actionable to most general. I finish with some thoughts on what to make of this unexpected proliferation of arguments. Primarily, I think it increases the importance of clarifying and debating the core ideas in AI safety.

1. *Maximisers are dangerous.* Superintelligent AGI will behave as if it's maximising the expectation of some utility function, since doing otherwise can be shown to be irrational. Yet we can't write down a utility function which precisely describes human values, and optimising very hard for any other function will lead to that AI rapidly seizing control (as a convergent instrumental subgoal) and building a future which contains very little of what we value (because of Goodhart's law and the complexity and fragility of values). We won't have a chance to notice and correct misalignment because an AI which has exceeded human level will increase its intelligence very quickly (either by recursive self-improvement or by scaling up its hardware), and then prevent us from modifying it or shutting it down.
    1. This was the main thesis advanced by Yudkowsky and Bostrom when founding the field of AI safety. Here I've tried to convey the original line of argument, although some parts of it have been strongly critiqued since then. In particular, Drexler and Shah have disputed the relevance of expected utility maximisation (the latter suggesting the concept of goal-directedness as a replacement), while Hanson and Christiano disagree that AI intelligence will increase in a very fast and discontinuous way.
    2. Most of the arguments in this post originate from or build on this one in some way. This is particularly true of the next two arguments - nevertheless, I think that there's enough of a shift in focus in each to warrant separate listings.
2. *The target loading problem.* Even if we knew exactly what we wanted a superintelligent agent to do, we don't currently know (even in theory) how to make an agent which actually tries to do that. In other words, if we were to create a superintelligent AGI before solving this problem, the goals we would

ascribe to that AGI (by taking the intentional stance towards it) would not be the ones we had intended to give it. As a motivating example, evolution selected humans for their genetic fitness, yet humans have goals which are very different from just spreading their genes. In a machine learning context, while we can specify a finite number of data points and their rewards, neural networks may then extrapolate from these rewards in non-humanlike ways.
    1. This is a more general version of the "inner optimiser problem", and I think it captures the main thrust of the latter while avoiding the difficulties of defining what actually counts as an "optimiser". I'm grateful to Nate Soares for explaining the distinction, and arguing for the importance of this problem.
3. *The prosaic alignment problem.* It is plausible that we build "prosaic AGI", which replicates human behaviour without requiring breakthroughs in our understanding of intelligence. Shortly after they reach human level (or possibly even before), such AIs will become the world's dominant economic actors. They will quickly come to control the most important corporations, earn most of the money, and wield enough political influence that we will be unable to coordinate to place limits on their use. Due to economic pressures, corporations or nations who slow down AI development and deployment in order to focus on aligning their AI more closely with their values will be outcompeted. As AIs exceed human-level intelligence, their decisions will become too complex for humans to understand or provide feedback on (unless we develop new techniques for doing so), and eventually we will no longer be able to correct the divergences between their values and ours. Thus the majority of the resources in the far future will be controlled by AIs which don't prioritise human values. This argument was explained in this blog post by Paul Christiano.
    1. More generally, aligning multiple agents with multiple humans is much harder than aligning one agent with one human, because value differences might lead to competition and conflict even between agents that are each fully aligned with some humans. (As my own speculation, it's also possible that having multiple agents would increase the difficulty of single-agent alignment - e.g. the question "what would humans want if I didn't manipulate them" would no longer track our values if we would counterfactually be manipulated by a different agent).
4. *The human safety problem.* This line of argument (which Wei Dai has recently highlighted) claims that no human is "safe" in the sense that giving them absolute power would produce good futures for humanity in the long term, and therefore that building AI which extrapolates and implements the values of even a very altruistic human is insufficient. A prosaic version of this argument emphasises the corrupting effect of power, and the fact that morality is deeply intertwined with social signalling - however, I think there's a stronger and more subtle version. In everyday life it makes sense to model humans as mostly rational agents pursuing their goals and values. However, this abstraction breaks down badly in more extreme cases (e.g. addictive superstimuli, unusual moral predicaments), implying that human values are somewhat incoherent. One such extreme case is running my brain for a billion years, after which it seems very likely that my values will have shifted or distorted radically, in a way that my original self wouldn't endorse. Yet if we want a good future, this is the process which we require to go well: a human (or a succession of humans) needs to maintain broadly acceptable and coherent values for astronomically long time periods.
    1. An obvious response is that we shouldn't entrust the future to one human, but rather to some group of humans following a set of decision-making procedures. However, I don't think any currently-known institution is

actually much safer than individuals over the sort of timeframes we're talking about. Presumably a committee of several individuals would have lower variance than just one, but as that committee grows you start running into well-known problems with democracy. And while democracy isn't a bad system, it seems unlikely to be robust on the timeframe of millennia or longer. (Alex Zhu has made the interesting argument that the problem of an individual maintaining coherent values is roughly isomorphic to the problem of a civilisation doing so, since both are complex systems composed of individual "modules" which often want different things.)
    2. While AGI amplifies the human safety problem, it may also help solve it if we can use it to decrease the value drift that would otherwise occur. Also, while it's possible that we need to solve this problem in conjunction with other AI safety problems, it might be postponable until after we've achieved civilisational stability.
    3. Note that I use "broadly acceptable values" rather than "our own values", because it's very unclear to me which types or extent of value evolution we should be okay with. Nevertheless, there are some values which we definitely find unacceptable (e.g. having a very narrow moral circle, or wanting your enemies to suffer as much as possible) and I'm not confident that we'll avoid drifting into them by default.
5. *Misuse and vulnerabilities*. These might be catastrophic even if AGI always carries out our intentions to the best of its ability:
    1. AI which is superhuman at science and engineering R&D will be able to invent very destructive weapons much faster than humans can. Humans may well be irrational or malicious enough to use such weapons even when doing so would lead to our extinction, especially if they're invented before we improve our global coordination mechanisms. It's also possible that we invent some technology which destroys us unexpectedly, either through unluckiness or carelessness. For more on the dangers from technological progress in general, see Bostrom's paper on the vulnerable world hypothesis.
    2. AI could be used to disrupt political structures, for example via unprecedentedly effective psychological manipulation. In an extreme case, it could be used to establish very stable totalitarianism, with automated surveillance and enforcement mechanisms ensuring an unshakeable monopoly on power for leaders.
    3. AI could be used for large-scale projects (e.g. climate engineering to prevent global warming, or managing the colonisation of the galaxy) without sufficient oversight or verification of robustness. Software or hardware bugs might then induce the AI to make unintentional yet catastrophic mistakes.
    4. People could use AIs to hack critical infrastructure (include the other AIs which manage aforementioned large-scale projects). In addition to exploiting standard security vulnerabilities, hackers might induce mistakes using adversarial examples or 'data poisoning'.
6. *Argument from large impacts*. Even if we're very uncertain about what AGI development and deployment will look like, it seems likely that AGI will have a very large impact on the world in general, and that further investigation into how to direct that impact could prove very valuable.
    1. Weak version: development of AGI will be at least as big an economic jump as the industrial revolution, and therefore affect the trajectory of the long-term future. See Ben Garfinkel's talk at EA Global London 2018. Ben noted that to consider work on AI safety important, we also need to believe the additional claim that there are feasible ways to positively influence the

long-term effects of AI development - something which may not have been true for the industrial revolution. (Personally my guess is that since AI development will happen more quickly than the industrial revolution, power will be more concentrated during the transition period, and so influencing its long-term effects will be more tractable.)
2. Strong version: development of AGI will make humans the second most intelligent species on the planet. Given that it was our intelligence which allowed us to control the world to the large extent that we do, we should expect that entities which are much more intelligent than us will end up controlling our future, unless there are reliable and feasible ways to prevent it. So far we have not discovered any.

What should we think about the fact that there are so many arguments for the same conclusion? As a general rule, the more arguments support a statement, the more likely it is to be true. However, I'm inclined to believe that quality matters much more than quantity - it's easy to make up weak arguments, but you only need one strong one to outweigh all of them. And this proliferation of arguments is (weak) evidence against their quality: if the conclusions of a field remain the same but the reasons given for holding those conclusions change, that's a warning sign for motivated cognition (especially when those beliefs are considered socially important). This problem is exacerbated by a lack of clarity about which assumptions and conclusions are shared between arguments, and which aren't.

On the other hand, superintelligent AGI is a very complicated topic, and so perhaps it's natural that there are many different lines of thought. One way to put this in perspective (which I credit to Beth Barnes) is to think about the arguments which might have been given for worrying about nuclear weapons, before they had been developed. Off the top of my head, there are at least four:

1. They might be used deliberately.
2. They might be set off accidentally.
3. They might cause a nuclear chain reaction much larger than anticipated.
4. They might destabilise politics, either domestically or internationally.

And there are probably more which would have been credible at the time, but which seem silly now due to hindsight bias. So if there'd been an active anti-nuclear movement in the 30's or early 40's, the motivations of its members might well have been as disparate as those of AI safety advocates today. Yet the overall concern would have been (and still is) totally valid and reasonable.

I think the main takeaway from this post is that the AI safety community as a whole is still confused about the very problem we are facing. The only way to dissolve this tangle is to have more communication and clarification of the fundamental ideas in AI safety, particularly in the form of writing which is made widely available. And while it would be great to have AI safety researchers explaining their perspectives more often, I think there is still a lot of explicatory work which can be done regardless of technical background. In addition to analysis of the arguments discussed in this post, I think it would be particularly useful to see more descriptions of deployment scenarios and corresponding threat models. It would also be valuable for research agendas to highlight which problem they are addressing, and the assumptions they require to succeed.

This post has benefited greatly from feedback from Rohin Shah, Alex Zhu, Beth Barnes, Adam Marblestone, Toby Ord, and the DeepMind safety team. All opinions are

my own.

What are the open problems in Human Rationality?

LessWrong has been around for 10+ years, CFAR's been at work for around 6, and I think there have been at least a few other groups or individuals working on what I think of as the "Human Rationality Project."

I'm interested, especially from people who have invested significant time in attempting to push the rationality project forward, what they consider the major open questions facing the field. (More details in this comment)

"What is the Rationality Project?"

I'd prefer to leave "Rationality Project" somewhat vague, but I'd roughly summarize it as "*the study of how to have optimal beliefs and make optimal decisions while running on human wetware.*"

If you have your own sense of what this means or should mean, feel free to use that in your answer. But some bits of context for a few possible avenues you could interpret this through:

Early LessWrong focused a lot of cognitive biases and how to account for them, as well as Bayesian epistemology.

CFAR (to my knowledge, roughly) started from a similar vantage point and eventually started moving in the direction of "how to do you figure out what you actually want and bring yourself into 'internal alignment' when you want multiple things, and/or different parts of you want different things and are working at cross purposes. It also looked a lot into Double Crux, as a tool to help people disagree more productively.

CFAR and Leverage both ended up exploring introspection as a tool.

Forecasting as a field has matured a bit. We have the Good Judgment project.

Behavioral Economics has begun to develop as a field.

I recently read "How to Measure Anything", and was somewhat struck at how it tackled prediction, calibration and determining key uncertainties in a fairly rigorous, professionalized fashion. I could imagine an alternate history of LessWrong that had emphasized this more strongly.

With this vague constellation of organizations and research areas, gesturing at an overall field...

...what are the big open questions the field of Human Rationality needs to answer, in order to help people have more accurate beliefs and/or make better decisions?

Announcement: AI alignment prize round 4 winners

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

We (Zvi Mowshowitz and Vladimir Slepnev) are happy to announce the results of the fourth round of the AI Alignment Prize, funded by Paul Christiano. From July 15 to December 31, 2018 we received 10 entries, and are awarding four prizes for a total of $20,000.

The winners

We are awarding two first prizes of $7,500 each. One of them goes to Alexander Turner for Penalizing Impact via Attainable Utility Preservation; the other goes to Abram Demski and Scott Garrabrant for the Embedded Agency sequence.

We are also awarding two second prizes of $2,500 each: to Ryan Carey for Addressing three problems with counterfactual corrigibility, and to Wei Dai for Three AI Safety Related Ideas and Two Neglected Problems in Human-AI Safety.

We will contact each winner by email to arrange transfer of money. Many thanks to everyone else who participated!

Moving on

This concludes the AI Alignment Prize for now. It has stimulated a lot of good work during its year-long run, but participation has been slowing down from round to round, and we don't think it's worth continuing in its current form.

Once again, we'd like to thank everyone who sent us articles! And special thanks to Ben and Oliver from the LW2.0 team for their enthusiasm and help.

Strategy is the Deconfusion of Action

Reading the New Research Directions update from MIRI, I was struck by the description of deconfusion:

> By deconfusion, I mean something like “making it so that you can think about a given topic without continuously accidentally spouting nonsense.”

I find this concept deeply impressive. I have also lately been considering the problem of strategy or lack thereof, so the idea popped up almost immediately: strategy is deconfusion in the action domain.

I'm going to draw on three sources for this post: the first is the aforementioned New Research Directions post from MIRI; the second is a paper from Parameters 46 Winter issue by Jeffrey W. Meiser, "Are Our Strategic Models Flawed? Ends+Ways+Means = (Bad) Strategy"; the third is an article from November 2012 in The Atlantic by Thomas E. Ricks, "General Failure." I recommend them all individually, but I won't assume you have read them.

Returning to the concept of deconfusion, it is easy to change the quoted section only a little to capture what I mean:

> By strategy, I mean something like "making it so that you can act towards a given end-state without continuously accidentally wasting effort."

I think this is an important connection to draw. There is a *lot* of information available on strategy: each military philosopher of note supports an entire corpus of commentary, and likewise for every conqueror; every war has lots of official and academic analysis done on its results; there is an absurd profusion of filtering the military and historical information through the lens of self-help or business-speak. It has all *very consistently* failed to guide the development and execution of good strategies, and I think deconfusion does a good job of pointing to why.

Failure to Notice Confusion and the Lykke Model

Meiser's paper is about the current norms in the US military and the Army in particular. The focus of the paper is the Lykke model and the ways in which it encourages bad strategy. It consists of the following:

> “Strategy equals ends (objectives toward which one strives) plus ways (courses of action) plus means (instruments by which some end can be achieved).”

The problem in practice, Meiser argues, is that strategy development is dominated by means-based planning. This is because the theory is from 1989, developed in reaction to the failures of the Vietnam War; the thinking went that if resource constraints were better taken into account, we could prevent "unrealistic strategies." Ends are treated as given and what people mostly do with planning is look at the means table, match them against the ends table, and then call it a day. The problem with the model is that it promotes a kind of plug-and-chug approach.

Meiser uses General Stanley McChrystal's plan for Afghanistan after he took command of that theater as an example:

> What emerges from journalistic accounts of the 2009 Obama administration strategy-making process is the observation that the entire discussion by civilian officials and military officers was about the number of troops, not strategy.
>
> . . .
>
> After repeated presidential requests for at least three distinct options, all Obama ever got was slight variations of the original ones. All options were based on the amount of resources being thrown at the problem.

The debate was actually moot. No one knew how to really use those resources, and the military did not notice their confusion about the problem. Of course, this is not the only shortfall.

Assuming Confusion Away

Returning to the MIRI post, this is given as an example of confused thinking about AI risk:

> People who are serious thinkers about the topic today, including my colleagues Eliezer and Anna, said things that today sound confused. (When I say "things that sound confused," I have in mind things like "isn't intelligence an incoherent concept," "but the economy's already superintelligent," "if a superhuman AI is smart enough that it could kill us, it'll also be smart enough to see that that isn't what the good thing to do is, so we'll be fine," "we're Turing-complete, so it's impossible to have something dangerously smarter than us, because Turing-complete computations can emulate anything," and "anyhow, we could just unplug it.")

From the Atlantic article, quoting a slide from a classified briefing:

> "What to Expect After Regime Change":
>
> Most tribesmen, including Sunni loyalists, will realize that their lives will be better once Saddam is gone for good. Reporting indicates a growing sense of fatalism, and accepting their fate, among Sunnis. There may be a small group of die hard supporters that [are] willing to rally in the regime's heartland near Tikrit—but they won't last long without support.

Comparing these two is only marginally appropriate; the first quote is from (at the time) amateurs who were deeply engaged with the problem they were thinking about, whereas the latter is from nominal experts who were negligently hand-waving the problem away. I say that both suggest confusion about how to consider the problems at hand. I go further and say that the latter is a more pernicious sort of confusion, because they assumed there never was any from the beginning.

Also from Ricks, regarding General Tommy Franks at the beginning of the Iraq War:

> In many ways, Franks is the representative general of the post-9/11 era. He concerned himself principally with tactical matters, refusing to think seriously about what would happen after his forces attacked. "I knew the President and Don Rumsfeld would back me up," he wrote in his memoir, "so I felt free to pass the message along to the bureaucracy beneath them: *You pay attention to the day* after *and I'll pay attention to the day* of." Franks fundamentally misunderstood

> generalship, which at its topmost levels must link military action to political results.

This did not improve as the war went on:

> Not long after the Anaconda battle, Franks spoke at the Naval War College, in Newport, Rhode Island. A student heard his talk and then posed the most basic but most important sort of question: What is the nature of the war you are fighting in Afghanistan? “That’s a great question for historians,” Franks said. He then went on to discuss how U.S. troops cleared cave complexes.

Nor was it a problem unique to Franks. Regarding his successor, Lieutenant General Ricardo Sanchez:

> Sanchez inherited no real war strategy from Franks or the Bush administration, and he did nothing to remedy that deficit. This lack of any coherent strategy manifested itself in the radically different approaches taken by different Army divisions in the war. Observers moving from one part of Iraq to another were often struck by the extent to which each division was fighting its own war, with its own assessment of the threat, its own solutions, and its own rules of engagement.

Eventually on the third try someone noticed that they were confused about what the Army was doing, in the person of General George Casey:

> He knew the Army needed to start operating differently in Iraq. He developed a formal campaign plan, something Sanchez had never done. More significant, he asked two counterinsurgency experts, Colonel Bill Hix and retired Lieutenant Colonel Kalev Sepp, to review the actions of individual units and make suggestions. Sepp, a Special Forces veteran of El Salvador with a doctorate from Harvard, reviewed the commander of every battalion, regiment, and brigade in Iraq and concluded that 20 percent of them understood how to properly conduct counterinsurgency operations, 60 percent were struggling to do so, and 20 percent were not interested in changing and were fighting conventionally, “oblivious to the inefficacy and counterproductivity of their operations.” In other words, a vast majority of U.S. units were not operating effectively.

But he, too, took the ends as both given and uncomplicated. While Baghdad was in the throes of civil war:

> Casey’s lack of awareness began to undercut his support at the top of the Bush administration. On August 17, 2006, during a video briefing to top national-security officials, he said he wanted to stick with his plan to turn Baghdad over to Iraqi security forces by the end of the year. Vice President Dick Cheney, watching from Wyoming, was troubled by that comment. “I respected General Casey, but I couldn’t see a basis for his optimism,” he wrote later.

From the beginning of the conflicts in both Afghanistan and Iraq, the possibility of being confused about the goals was not seriously considered, because the military assumed the problem away.

Theory of Success and Re-enter Deconfusion

Current strategy norms don't admit the idea of confusion, which is a problem. In Meiser's paper he offers a different definition of strategy which he hopes will promote "creative and critical thinking," which I have taken the liberty of interpreting as

addressing the confusion problem. Pleasingly he moves into terms and concepts we are familiar with (emphasis mine):

> The two definitions that come closest to articulating a distinctive meaning for strategy are offered by Barry Posen and Eliot Cohen. Posen defines grand strategy as “a state’s theory about how it can best ‘cause’ security for itself.” Cohen defines strategy as a “theory of victory.” The key insight by Posen and Cohen is the inclusion of the term theory. If we define theories as “statements predicting which actions will lead to what results—and why,” we can move toward a better definition of strategy that is general, but not too inclusive, and captures the essence of the concept.
>
> If we use the Posen-Cohen approach with a more general definition of purpose, we arrive at a sufficient working definition: strategy is a theory of success. This creates the expectation that **anything called a strategy will be a causal explanation of how a given action or set of actions will cause success**. Most strategies will include multiple intervening variables and conditions. Defining strategy as a theory of success encourages creative thinking while keeping the strategist rooted in the process of causal analysis; it brings assumptions to light and forces strategists to clarify exactly how they plan to cause the desired end state to occur.

This puts strategy on the same type of conceptual ground that motivated deconfusion in the first place; it is the primary reason I see deconfusion being so valuable. A secondary reason I see deconfusion as being valuable is shifting from the current paradigm. Currently formal strategic methods don't account for confusion, and lazy or negligent approaches to those methods can make it impossible to resolve. Viewing deconfusion as fundamental means that my assumption is that *I am confused.*

Combat vs Nurture & Meta-Contrarianism

My initial reaction to Combat vs Nurture was to think "I already wrote about that!" and "but there should be three clusters, not two". However, looking at my old posts, I see that my thinking has shifted since I wrote them, and I don't describe the three clusters in quite the way I currently would. So, here is how I think about it:

- **"Face Culture" / "Playing Team":** When people offer ideas, their ego/reputation is on the line. It is therefore important to "recognize the value of every contribution" -- accepting or rejecting an idea has a strong undercurrent of accepting or rejecting the person offering the idea. This sometimes makes rational decision impossible; the value of team members is greater than the value of the specific decision, so incorporating input from a team member can be a higher priority than making the best decision. Much of the time, bad ideas *can* be discarded, but it involves a dance of "due consideration": an idea may be entertained for longer than necessary in order to signal strongly that the team valued the input, and downsides may be downplayed to avoid making anyone look stupid. Eventually you want someone on the team to point out a decisive downside in order to reject a bad idea, but ideally you coax this out of the person who originated the idea. (I called this "face culture", but I have heard people call it "playing team".)
- **Intellectual Debate:** It is assumed that engaging with an idea involves arguing against it. Approving an idea doesn't necessarily signal anything bad, but unlike face culture, arguing against an idea doesn't signal anything bad either. Arguing against someone (genuinely or in a devil's-advocate position) shows that you find the idea interesting and worth engaging with. Concepts like burden of proof are often applied; one tends to operate as if there were an objective standard of truth which both sides of the debate are accountable to. This warps the epistemic standards in a variety of ways, for example, making it unacceptable to bring raw intuitions to the table without putting them in justifiable terms (even if a raw intuition is your honest reason). However, if you have to choose between face culture and intellectual debate culture, intellectual debate is far better for making intellectual progress.
- **Mutual Curiosity & Exploration:** I called this level "intellectual honesty" in my old post. This level is closer to the spirit of double crux or circling. In this type of conversation, there may still be some "sides" to debate, but everyone is on the side of the truth; there is no need for someone to take one side or the other, except to the extent that they hold some intuitions which haven't been conveyed to others yet. In other words, it is more natural to weave around offering supporting/contrary evidence for various possibilities, instead of sticking to one side and defending it while attacking others. It is also more natural for there to be more than two possibilities on the table (or more possibilities than people in the conversation). People don't need to have any initial disagreement in order to have this kind of conversation.

Whereas Ruby's Combat vs Nurture post put the two cultures on a roughly even footing, I've obviously created a hierarchy here. But, the hierarchy swings between the two poles of combat and nurture. Ruby mentioned that there's a contrarian aspect to intellectual debate: the bluntness manages to be a countersignal to the more mainstream niceness signal, so that getting blunt responses actually signals social

acceptance. Yet, Ruby also mentions that the culture amongst bay area rationalists is primarily nurture culture, seemingly aligning with the mainstream rather than the contrarian combat culture. I explain this by offering my three-layer cake above, with mutual curiosity and exploration being the meta-contrarian position. Although it can be seen as a return to nurture culture, it still *significantly differs* from what I call face culture.

I've said this before, and I'll say this again: placing these conversation cultures in a hierarchy from worse to better *does not mean that you should frown on "lower" strategies*. **It is very important to meet a conversation at the level at which it occurs**, respecting the games of face culture if they're being played. You can try to gently move a conversation in a positive direction, but a big part of the point of my original post on this stuff was to say that the underlying cause of different conversational practices is the *level of intellectual trust* present. Face culture is a way to manage conversations where people lack common knowledge of trust (and perhaps lack actual trust), so must signal carefully. Intellectual debate requires a level of safety such that you don't think an argument is a personal attack. Yet, at the same time, intellectual debate is a way of managing a discussion in which you can't trust people to be detached from their own ideas; you expect people to be biased in favor of what they've proposed, so you embrace that dynamic and construct a format where intellectual progress can happen anyway. The level of mutual curiosity and exploration can only be reached when there is trust that everyone has some ability to get past that bias. (Actually, double crux seems like a bridge between intellectual debate and mutual exploration, since it still leans heavily on the idea of people taking sides.)

Having established a meta-contrarian hierarchy, we can extend the idea further. This stretches things a bit, and I'm less confident that the five levels which follow line up with reality as well as the three I give above, but it seems worth mentioning:

- **0. Open Verbal Combat:** This is the "lower" level which face culture is a reaction to. Here, everyone's ego is out in the open. There is still a veneer of plausible deniability around intellectual honesty: arguments would be meaningless if no one respected the truth at all and only argued what was convenient to them in the moment. However, at this level, that's almost exclusively what's happening. Even in cases where it looks like arguments are being respected for their undeniable force, there's a lot of status dynamics in play; people are reacting to who they can expect to be on their side, and logic only has force as a coordinating signal.
- **1. Face Culture.**
- **2. Intellectual Debate.**
- **3. Mutual Curiosity.**
- **4. Exchanging Gears:** Once everyone has a common framework of mutual curiosity, in which exchanging intuitions is acceptable and valued rather than steamrolled by attempts at objectivity, then a further evolution is possible, which involves a slight shift back towards combat culture. At this level, you don't even worry very much about deciding on the truth of things. The focus is on exchanging possible models; you trust that everyone will go and observe the world later, and update in favor of the best models over a long period of time. Articulating and understanding models is the bottleneck, so it deserves most of the attention. I think this is what Ben Pace describes in Share Models, Not Beliefs. However, this shift is smaller than the shifts between levels below this one (at least, in terms of what I currently understand).

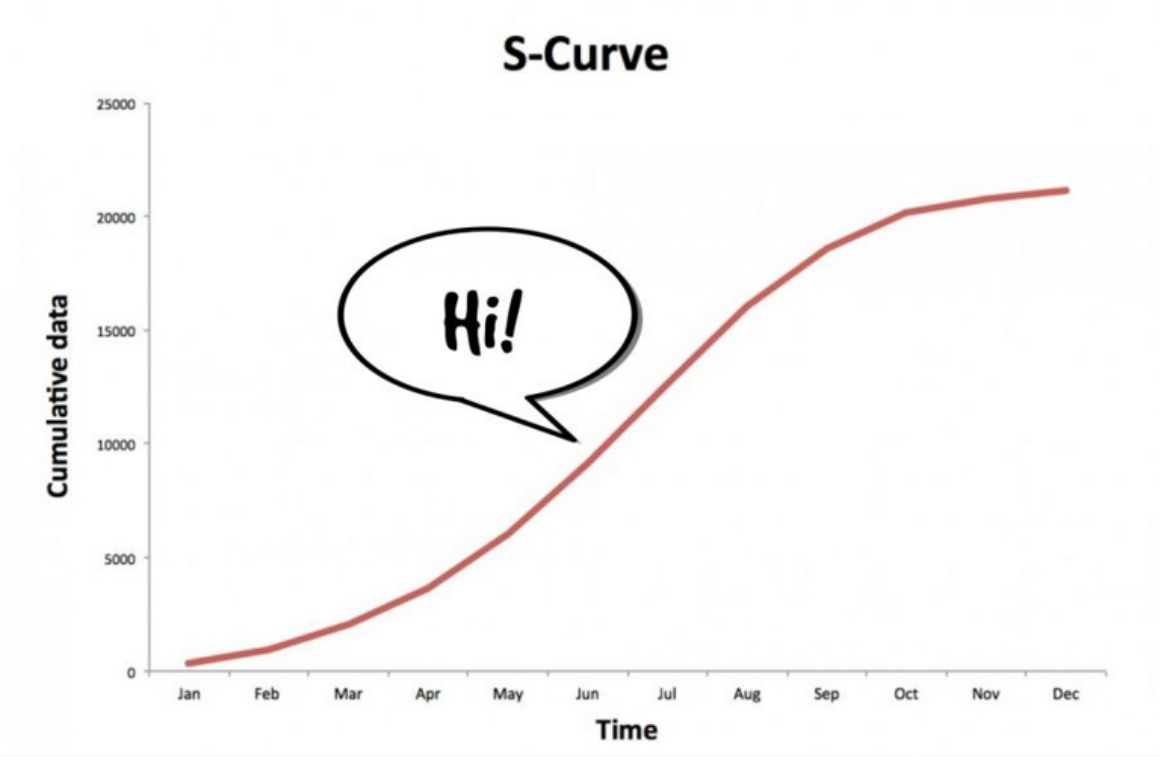
Again: *the biggest take-away from this should be that you want to* ***meet a conversation at the level at which it is occurring***. If you are used to one particular culture, you are very likely to be blind to what's going on in conversations following a different culture, and get frustrated or frustrate others. Read Surviving a Philosopher-Attack if you haven't, and keep in mind that responding from combat culture when someone is used to nurture culture can make people cry and never want to speak with you ever again.

S-Curves for Trend Forecasting

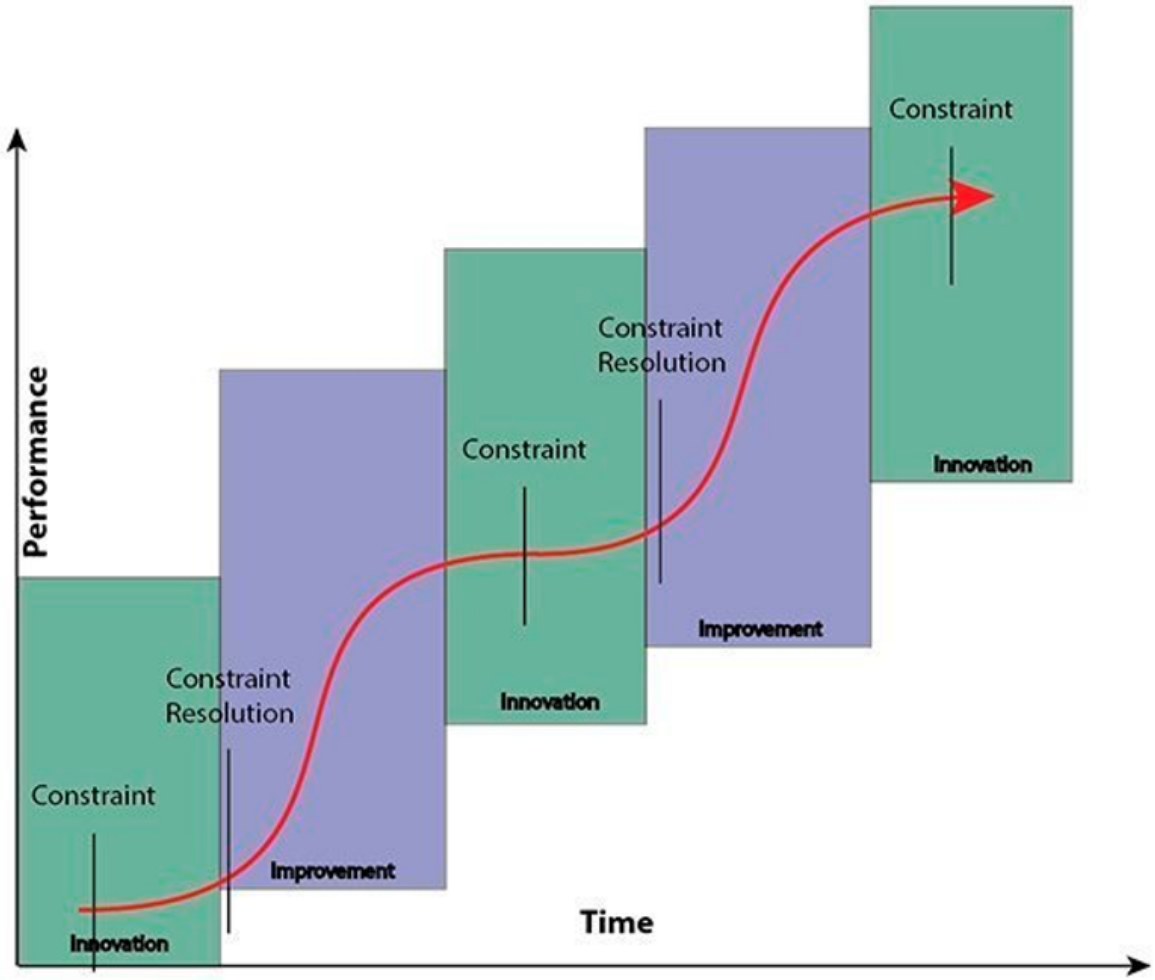
Epistemic Status: Innovation research and business research is notoriously low quality, and so all the ideas here should be viewed through that lense. What's impressive about the S-curve and evolution trends literature is how remarkably self-consistent it is using a wide variety of research methods. Whether Simon Wardley analyzing news article about different technologies, Clayton Christensen doing case studies of a specific industry, or Carlotta-Perez taking a historical approach of tracking different technologies, the same S-curve pattern and evolution trends seem to show up. This too should be taken into account when evaluating these ideas.

Basics

This is an S-curve.



The S-curve is a fundamental pattern that exists in many systems that have positive feedback loops and constraints. The curve speeds up due to the positive feedback loop, then slows down due to the constraints.



When the constraint is broken, the positive feedback loop ramps back up, until it hits another constraint.

Recommended Resource: Invisible Asymptotes , which gives a visceral feel for this process of positive feedback and constraints

Common Mistake: Confusing S-Curves With Exponential Growth

Sometimes, people get confused and call S-curves exponential growth. This isn't necessarily wrong but it can confuse their thinking. They forget that constraints exist and think that there will be exponential growth forever. When slowdowns happen, they think that it's the end of the growth - instead of considering that it may simply be another constraint and the start of another S-Curve. Knowledge of overlapping S-Curves can help you model these situations in a more sophisticated way.

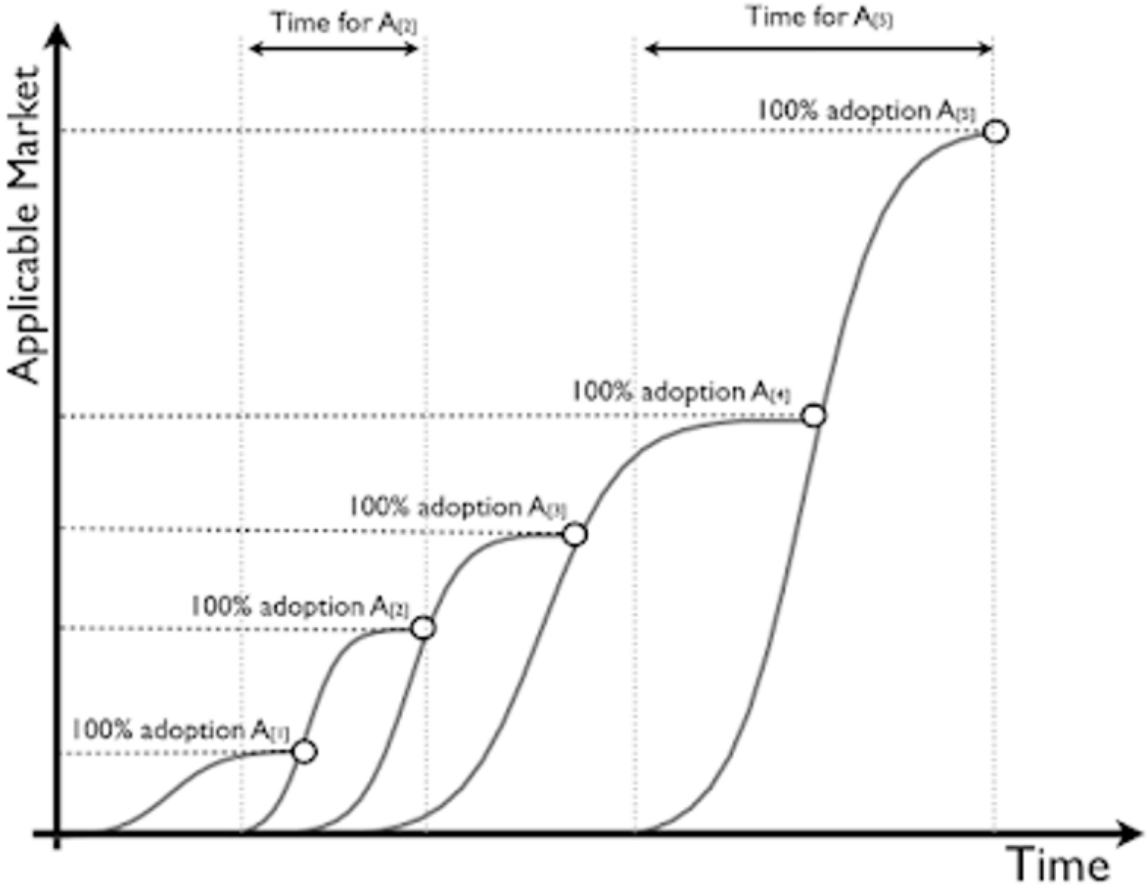
Diffusion S-Curves

The S-curve pattern is quite common in the spread of ideas, practices, and technologies, although it rarely looks quite as pretty. The example below shows "diffusion s-curves" - How a technology spreads through a population (in this case US households

The positive feedback loop in this case is word of mouth, and the constraints represent fundamental barriers to certain market segments or growth such as simplicity, usability, scalability, price, etc.

This creates smaller s-curves around adoption among specific market segments, and larger s-curves that represent the overall market penetration of the idea, practice, or technology.

Recommended Resource: [Wikipedia on Diffusion of Innovation]

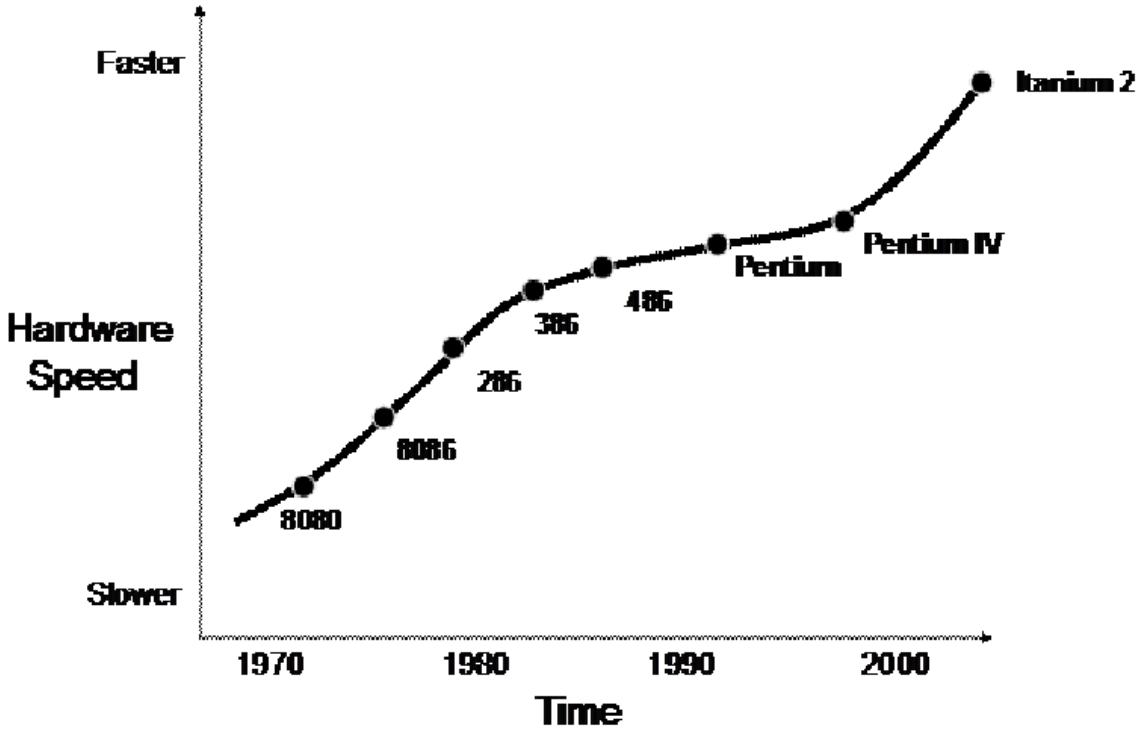


Evolution S-Curves

In addition to Diffusion S-curves in technology, ideas, and practices, there are Evolution S-Curves. These represent the increase in the traits of these ideas that make them usable in more situations and desirable for more people. When you break through a constraint in one of these properties through innovation, this can often coincide with "unlocking" a new diffusion curve by opening up a new market that wouldn't previously have used your technology or idea.

In this case the positive feedback loop is the increased understanding and expertise that comes from diffusion of a new innovation in your idea or technology, and the constraint represents fundamental assumptions in the idea, practice, or technology that must be changed through another innovation to make the idea, practice, or technology more desirable.

In the example below the desirable property is hardware speed. Fundamental leaps are made to break through a speed constraint, and then iterated on through the positive feedback loop of information and expertise increasing from adoption. This hits diminishing returns as the new innovation is optimized, and then a new fundamental innovation is needed to overcome the next constraint.



Recommended Resource: [Open University on Evolution S-Curves]

Common Mistake: Confusing Diffusion S-Curves with Evolution S-Curves

Sometimes, I see people make the mistake of assuming that evolution and diffusion s-curves follow the same cycle. Most often, the mistake made here is assuming that when a particular innovation has saturated a certain market, that also means it has "reached its final form" and has no more evolving to do.

There is a related truth - often, an innovation becoming more diffuse will drive innovation as new use cases become apparent. And vice versa, often new innovations will open a new market up by creating use cases that were previously impossible.

However, the two types of curves are driven by two different feedback loops and two different constraints. There's no reason to expect that they will follow each other, and no reason to expect that one curve leveling off will cause the other curve to level off.

S-Curves Patterns

S-curves become quite useful when paired with an understanding of evolutionary patterns. They can allow you to see in a broad sense what's coming next for an idea, practice or technology. They can prevent surprises and give you a tool to stay ahead of changes.

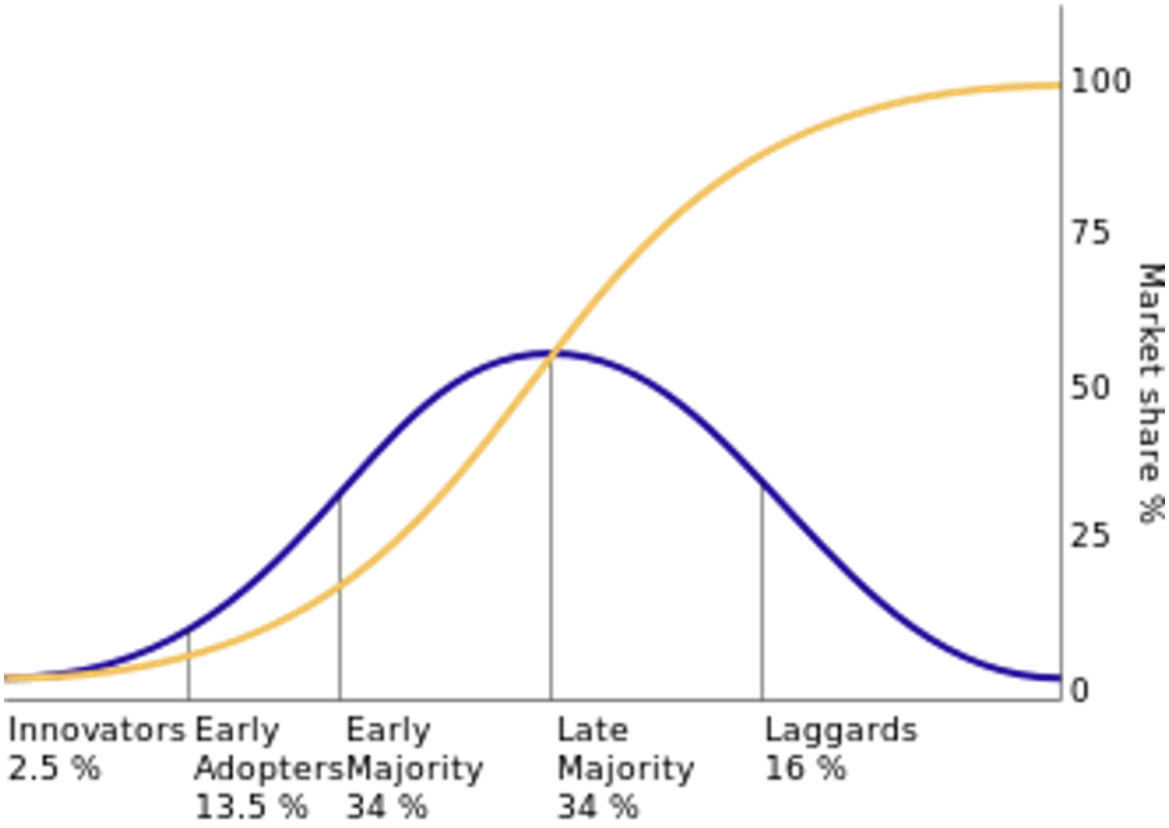
There are patterns that exist for both diffusion and evolution S-curves.

Diffusion Patterns

Diffusion patterns describe common themes that happen as trends diffuse through a population. They apply on the micro-level to individual population-segments, and on a macro-level to the overall population.

Diffusion of Innovation

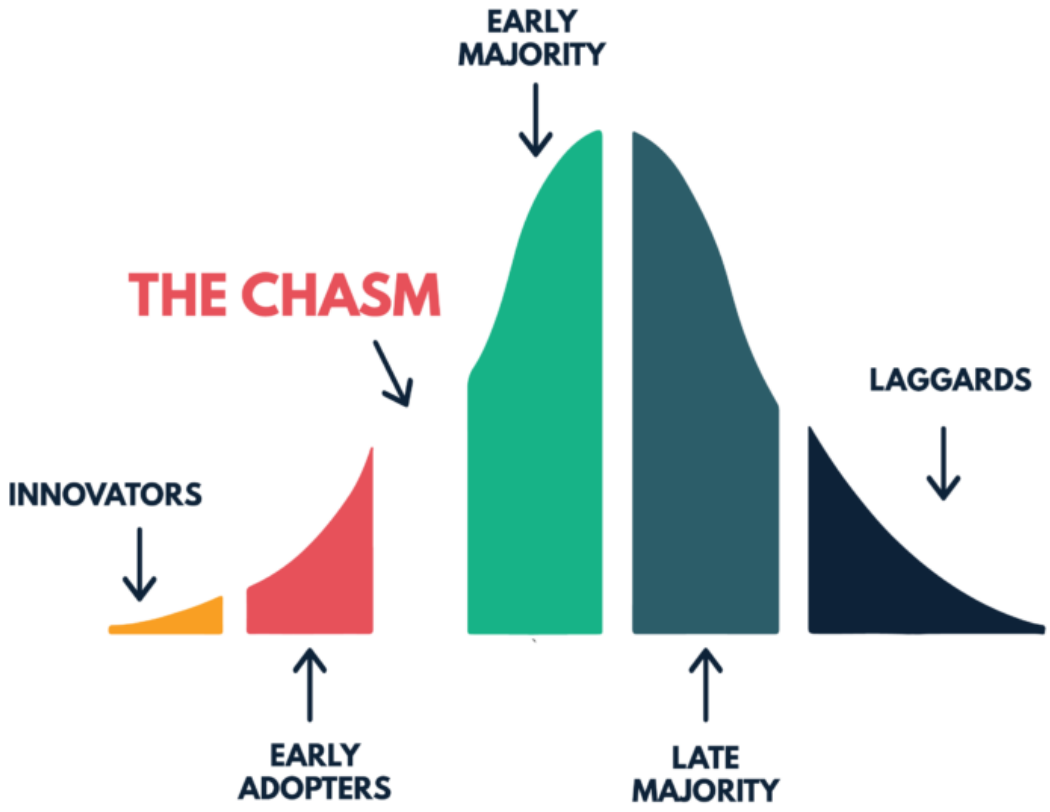
The diffusion of innovation describes 5 separate stages of a diffusion curve: Innovators, Early Adopters ,Early Majority, Late Majority, and Laggards. By understanding the traits of each of these groups, you can get a broad idea of what to expect, and how to slow or speed up adoption.



Recommended Resource: Diffusion of Innovations book by Everett Rogers

The Chasm

The Chasm describes a common constraint that occurs in a market segment between "early adopters" - who are willing to put up with a lot, and "early majority", who expect a lot. There is often a number of evolutionary constraints that must be broken through to bridge this single diffusion constraint and many new ideas, practices, and technologies get stuck in the chasm for that reason.



Recommended Resource: [Crossing the Chasm book by Geoffrey Moore]

Common Mistake: Assuming a Technology is Irrelevant Because it's Only Useful for a Small Group

A common mistake that I see is assuming a technology won't have a broader relevance, and using as evidence that it's only used by a small group of relatively abnormal people.

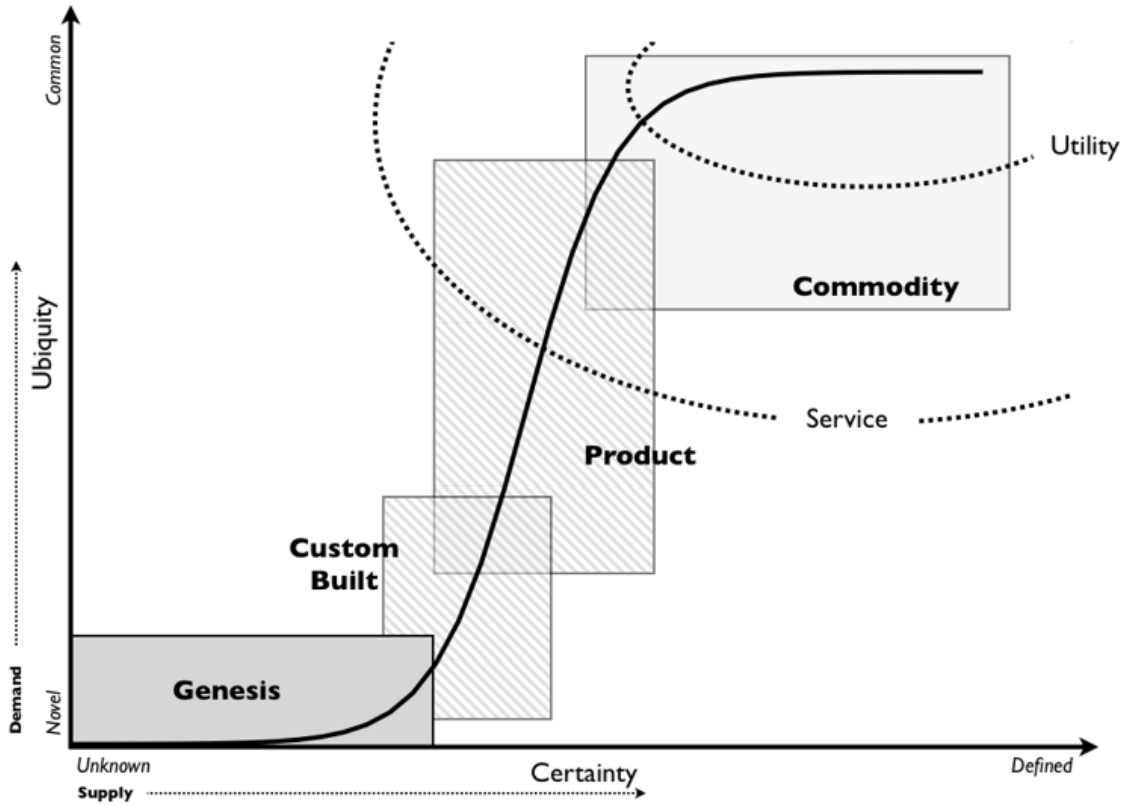
Now, what is true is that not all technologies eventually get adopted by everybody, some stay relatively niche. But it's not very good Bayesian evidence to say that because a technology is used by a small group of weird people, it will not have a broader impact. These diffusion patterns tell us that in fact that MOST technologies that eventually get widespread adoption go through this phase.

Furthermore, they tell us that many of those technologies often get stuck for a while at this early stage because of the Chasm. So even if a technology has staid at this tage for a while (e.g. cryptocurrency), it's still very little evidence towards that technology not being lifechanging in the future. (In contrast, a technology stalling for a long time at some point past the chasm is better evidence that it may have reached saturation)

Evolution Patterns

Evolution patterns describe common ways that innovations evolve over time to become increasingly desirable. They apply on the micro-level to individual innovations within a trend, and on a macro-level to the evolution of trend as a whole.

Wardley Evolution



Innovations tend to go through four stages - the initial prototype, custom built versions, productized versions that compete, than comoditized versions that are all basically the same. By understanding where you are, you can understand the type of competition likely to happen, the types of processes likely to yield improvements, and large changes that will be needed to stick with the market.

Recommended Resource: [Learn Wardley Mapping- Free Resource from Ben Mosior]

Common Mistake: Not reasoning about likely changes in how the market will be structured.

A common mistake I see when people reason about the future of e.g. Machine Learning, is that they reason as if the current economic style (how people make money from machine learning) will continue the way it has been.

What Wardley Evolution tells us is rather that it's very frequent for the way a market charges for and makes money with a particular innovation changes, and that change tends to fairly predictable.

For instance, I've seen analysis of Machine learning that assumes it will continue to be productized (which leads to very different dynamics in terms of competitive landscape and strategy between different AI vendors), rather than recognizing that it will eventually be commoditized and become a utility.

Simplicity - Complexity - Simplicity

Innovations tend to start out relatively simple as a new approach to a problem. They become increasingly complex to cover more use cases and be more robust, and then become simple again as refinements are made and they're distilled to their essence.

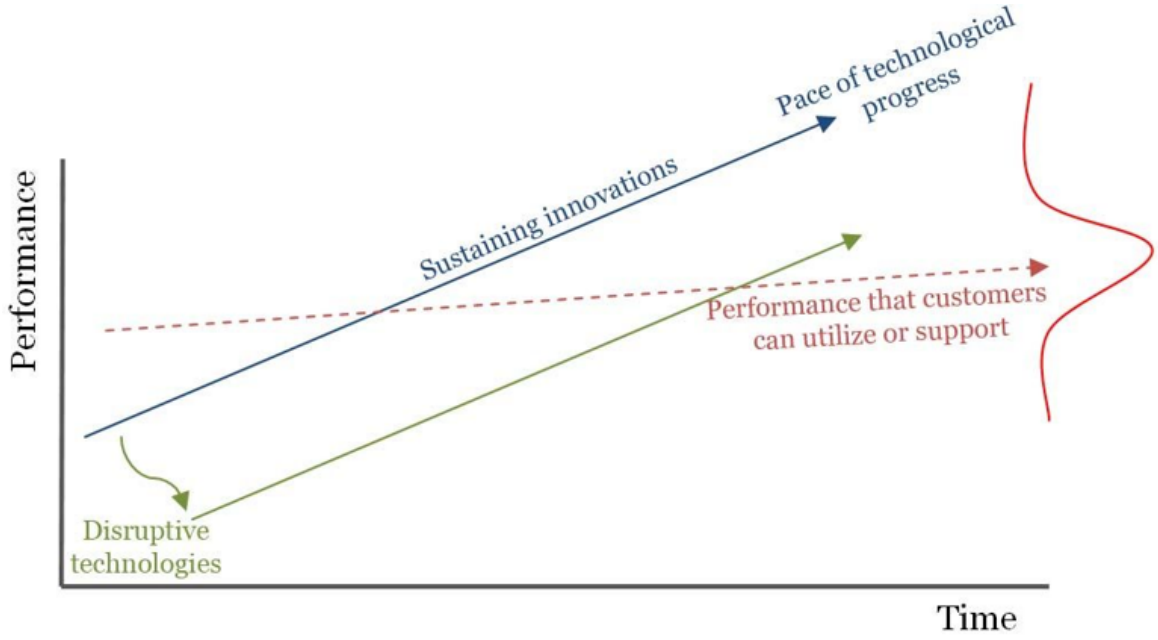
Recommended Resource: [TRIZ for Dummies book by Lilly Haines-Gadd]

Common Mistake: Assuming a Particular Innovation is a Dead End Because It's Gotten Too Complex

One mistake I see pretty frequently is people describing a particular innovation, and saying "well, we've added more and more complexity to this and it's gotten increasingly minimal returns so I expect there too not be too much more innovation in this area.

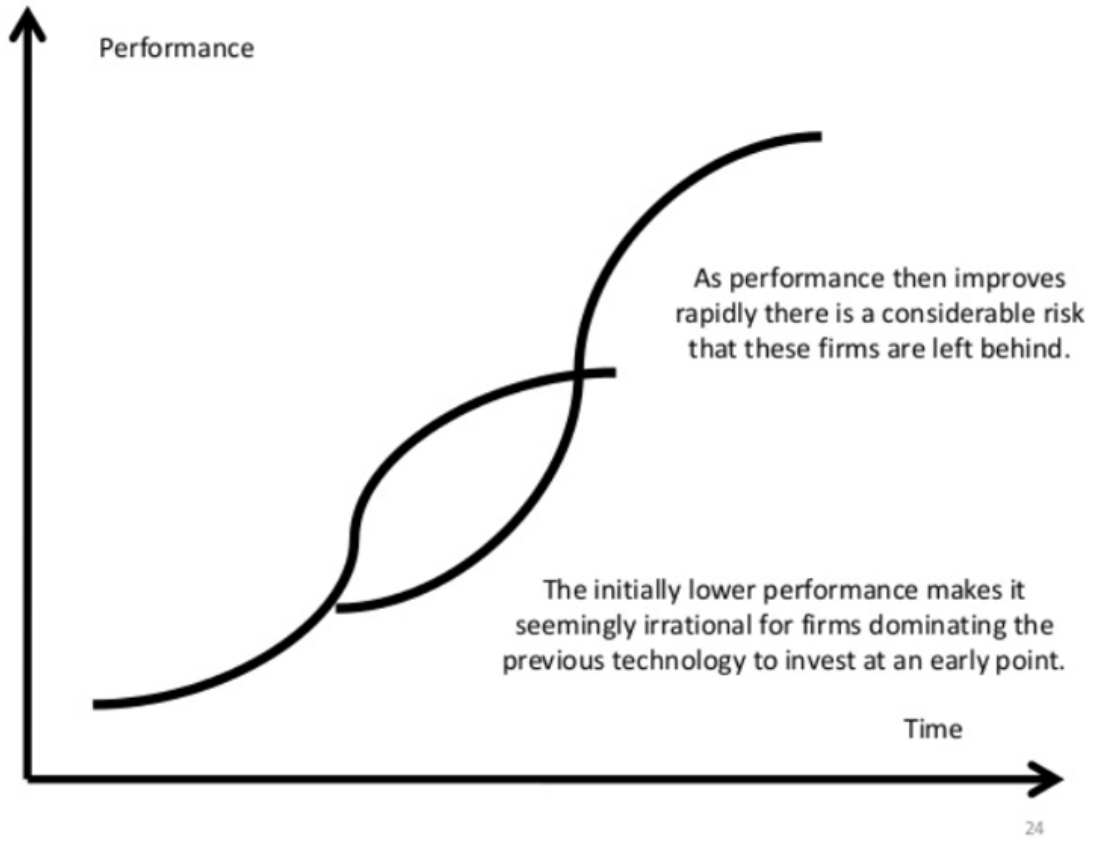
This can be true, but only if there are other indicators that this is already at the end of the innovation curve. Oftentimes, what's actually happened is that it's near the midpoint of it's innovation curve, and the next innovations will be around compressing/simplifying all the things that have been added. This simplification process then allows the innovation to be used a component to build further innovations off of, as it's simple enough to be commoditized.

Disruptive Innovation



Sometimes, innovations overshoot the mainstream populations needs on a particular dimension in order to be powerful for a particularly lucrative part of the population. In this case, these innovations or often overtaken by subsequent innovations that lower the performance on that dimension in order to raise it on other dimensions (example: Lower flexibility of a software product but raise the simplicity), these innovations can then "disrupt" the original innovation.

From the perspective a current innovation, the disruptive innovation appears to start below it in the s-curve, but it's able to gain adoption because the particular performance feature of that innovation is already higher than the market needs, and the new product competes on a different performance feature that is not even a target of.



Recommended Resource: [The Innovator's Dillema - Book by Clayton Christensen]

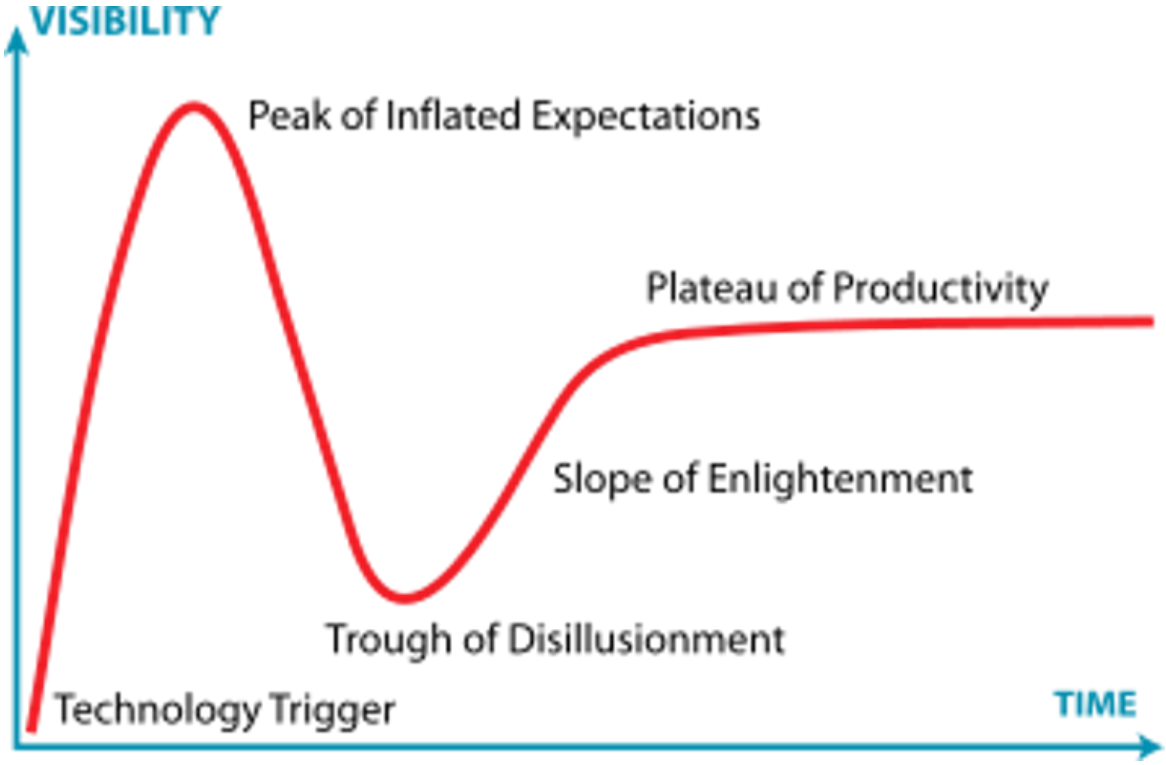
Common Mistake: Assuming a Particular Player will Win Because They're Big and Have Lots of Resources

One understandable assumption to make is that big players with more resources will always win. This isn't necessarily a bad assumption to make - disruptive innovations are much rarer than sustaining innovations.

However, having the disruptive innovation model can help you not make the mistake of just assuming that there's nothing that can topple the current champ - it gives you a clear model of exactly how this happens, and may even point out industries or areas where you're more likely to see this disruption take place.

Gartner Hype Cycle

The Gartner Hype Cycle describes a particular way that the media over-inflates people's expectations of new innovations in comparison to how evolved they actually are for a particular market segment's needs.

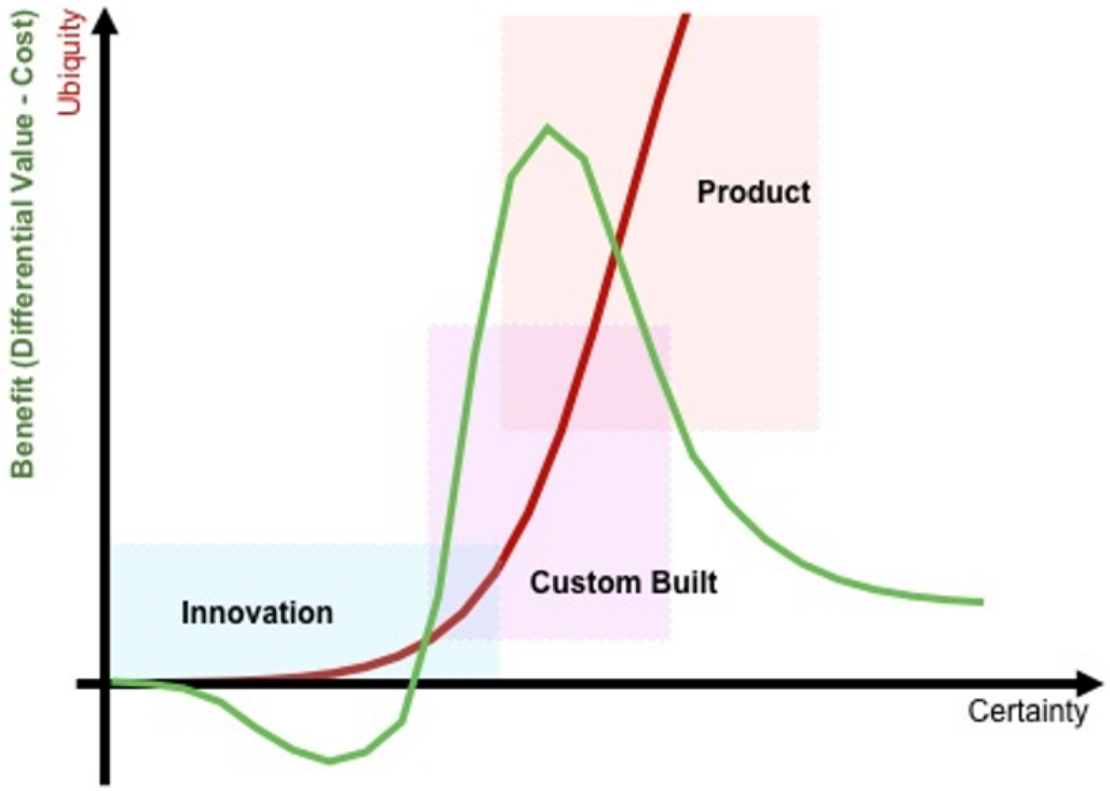


Recommended Resource: [Mastering the Hype Cycle - Book by Jackie Fenn]() (Disclaimer: Haven't read this one, only aware of the Gartner Hype Cycle in passing)

Common Mistake: Discounting a Particular Technology Because it Was Overhyped in the Past

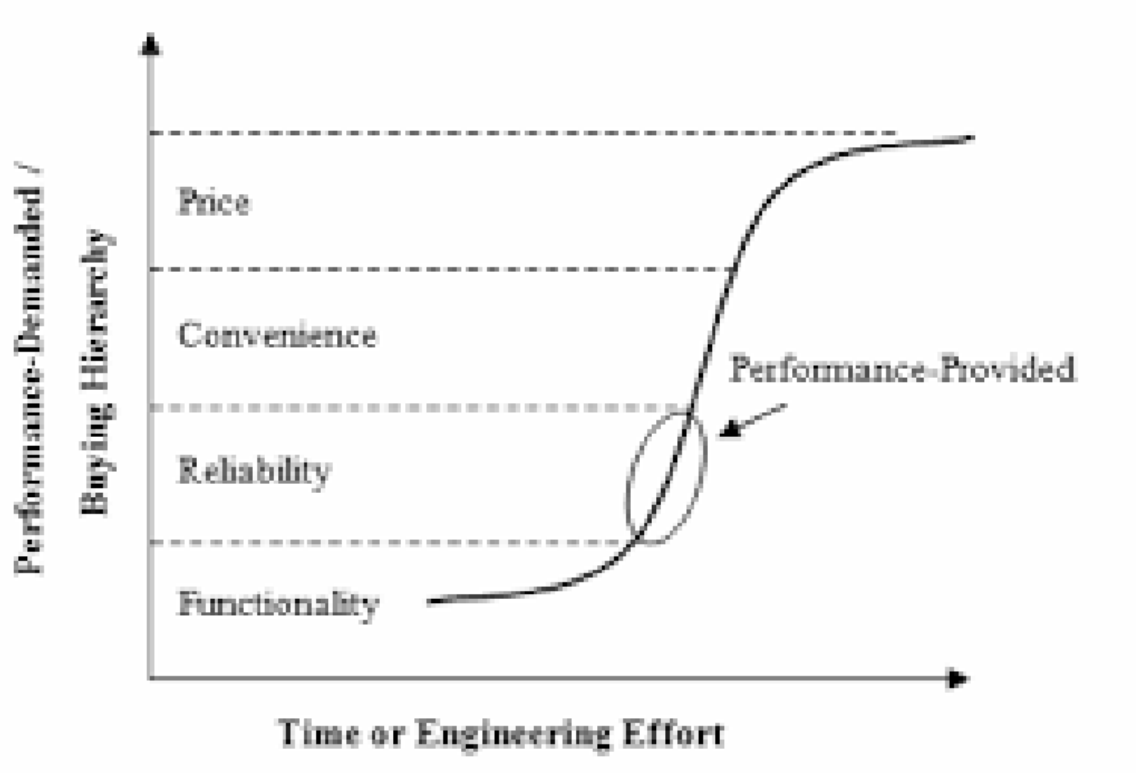
I've frequently seen arguments of the form - "Oh, you think this technology will have a massive impact? That's what they were saying a couple years ago and they massively overpromised.

Like other patterns, this is not saying that there aren't technologies that are massively overhyped and don't pan out. However, knowledge of the Gartner Hype cycle an show you that almost all popular technologies were once overhyped, so the argument of "this technology was overhyped in the past" isn't very good evidence of how transformative it will be. Rather, you'll want to map it against an evolution S-curve to see how overhyped you expect it to be relative to it's current level of evolution.



Windermere Buying Hierarchy

The Windermere Buying Hierarchy describes four different improvement focuses that an innovation optimizes over time. First, it's trying to solve for functionality, then reliability, then convenience, and finally price. This loosely maps to the stages of Wardley Evolution.



Recommended Resource: Haven't found a good one, learned about it through Clayton Christensen's work.

Common Mistake: Using Reliability, Convenience or Price as a Reason an Innovation Won't be Successful

You know the drill by now... it's not that reliability, convenience, or price are never reasons that a technology fails. But you'll want to map these against the evolution S-curves. It's common to see arguments about a technology not being viable because it's too expensive, when the S-curve is still WAYY at the early stage and we wouldn't even have expected the market to start thinknig about price optimization yet.

Only if the market has already reached that point in the S-curve and optimized that trait as much as it could, should you use this as a viable reason why you don't expect the technology to spread further.

Conclusion

S-curves and s-curve patterns are a useful tool for quickly analyzing systems, particularly when looking at diffusion of trends and evolution of innovations. They can heuristically identify solutions and probabilities that would otherwise be quite time consuming to figure out using something like a full system or functional analysis.

Hopefully you find this tool useful in your quest to understand all the things.

Megaproject management

Megaproject management is a new-ish subfield of project management. Originally considered to be the special case of project management where the budgets were enormous (billions of dollars), it is developing into a separate specialization because of the high complexity and tradition of failure among such projects. The driving force behind treating it as a separate field appears to be Bent Flyvbjerg, previously known around here for Reference Class Forecasting as the first person to develop an applied procedure. That procedure was motivated by megaprojects.

I will make a summary of the paper "What you should know about megaprojects, and why: an overview" from 2014. For casual reading, there is an article about it from the New Yorker here.

History

Megaprojects got their name from the association of mega with big, so think mega-city rather than mega-joule. It did match the unit prefix in the beginning however, as such projects were mostly dams, bridges, or very large buildings in the early 20th century.

The next shift upward took place with the Manhattan Project and then the Apollo program, which are also frequently drawn on as positive examples. The term 'megaproject' picked up steam in the 1970s, at the same time project costs crossed over into the billions.

Currently project costs of 50-100 billion are common, with even larger projects less common but not rare. If you were to view certain things which need dedicated management as a project, like the stimulus packages from 2008 or US defense procurement, then we have crossed over into the trillions and are entering a 'tera era' of megaprojects.

Ignoring these special cases, but counting infrastructure and industries where billion dollar projects are common, megaprojects account for ~8% of global GDP.

Four Sublimes

These are four reasons which drive the popularity of megaprojects. They are kind of a group bias for each type of stakeholder. They are:

- **Technological sublime:** because engineers and technologists love making the newest/tallest/fastest things.
- **Political sublime:** because politicians love being able to associate with huge projects and the publicity that comes with them.
- **Economic sublime:** because unions, contractors, and business people love all the jobs and fees.
- **Aesthetic sublime:** because designers love making beautiful things, and the public loves to adopt big beautiful things as distinctive for their city/country.

Predictably with biases, there are side effects:

> The following characteristics of megaprojects are typically overlooked or glossed over when the four sublimes are at play and the megaproject format is chosen for

delivery of large-scale ventures:

1. Megaprojects are inherently risky due to long planning horizons and complex interfaces (Flyvbjerg, 2006).

2. Often projects are led by planners and managers without deep domain experience who keep changing throughout the long project cycles that apply to megaprojects, leaving leadership weak.

3. Decision-making, planning, and management are typically multi-actor processes involving multiple stakeholders, public and private, with conflicting interests (Aaltonen and Kujala, 2010).

4. Technology and designs are often non-standard, leading to "uniqueness bias" amongst planners and managers, who tend to see their projects as singular, which impedes learning from other projects. 3

5. Frequently there is overcommitment to a certain project concept at an early stage, resulting in “lock-in” or “capture,” leaving alternatives analysis weak or absent, and leading to escalated commitment in later stages. "Fail fast" does not apply; "fail slow" does (Cantarelli et al., 2010; Ross and Staw, 1993; Drummond, 1998).

6. Due to the large sums of money involved, principal-agent problems and rent-seeking behavior are common, as is optimism bias (Eisenhardt, 1989; Stiglitz, 1989; Flyvbjerg el al., 2009).

7. The project scope or ambition level will typically change significantly over time.

8. Delivery is a high-risk, stochastic activity, with overexposure to so-called "black swans," i.e., extreme events with massively negative outcomes (Taleb, 2010). Managers tend to ignore this, treating projects as if they exist largely in a deterministic Newtonian world of cause, effect, and control.

9. Statistical evidence shows that such complexity and unplanned events are often unaccounted for, leaving budget and time contingencies inadequate.

10. As a consequence, misinformation about costs, schedules, benefits, and risks is the norm throughout project development and decision-making. The result is cost overruns, delays, and benefit shortfalls that undermine project viability during project implementation and operations.

The Iron Law of Megaprojects

- Over time.
- Over budget.
- Under utilized.

These aren't little, either: cost overruns of 1.5x are common, in bad cases they can run more than 10x, and 90% of projects have them; it is common for projects to have 0.5x *or less* utilization once complete. This holds for the public and private sectors, and also across countries, so things like excessive regulation or corruption aren't good explanations.

They start off badly, but they *do* still manage to get completed, which is due to...

Break-Fix Model

Since management of megaprojects doesn't know what they are doing or don't have the incentives to care, inevitably something breaks. Then additional time and money are spent to fix what broke, or the conditions of the project are renegotiated, and it limps along to the next break. This process continues until the project is finished.

If it is so terrible and we know it is terrible, why do we do it this way?

Hirschman's Hiding Hand

Because a lot of important stakeholders don't know how terrible it is. From Willie Brown, former mayor of San Francisco:

> "News that the Transbay Terminal is something like $300 million over budget should not come as a shock to anyone. We always knew the initial estimate was way under the real cost. Just like we never had a real cost for the [San Francisco] Central Subway or the [San Francisco-Oakland] Bay Bridge or any other massive construction project. So get off it. In the world of civic projects, the first budget is really just a down payment. If people knew the real cost from the start, nothing would ever be approved. The idea is to get going. Start digging a hole and make it so big, there's no alternative to coming up with the money to fill it in."

Nor are they without justification, for arguments have been made that support it. The first argument is exactly as Willie made it: if we knew how difficult large projects were, we would never build them.

For the second, note the title of the section is *hiding*, not *hidden*. This argument was made by Albert O. Hirschman on the basis of earlier work by J.E. Sawyer, and it says that there is an error in both the estimations of costs, and in the estimation of benefits, and this error should roughly cancel out. The problem is that Sawyer's work just pointed out that this was possible based on a picked sample of 5 or so. Hirschman then generalized it into a "Law of the Hiding Hand" and thereby legitimated lying to ourselves.

Alas it is bunk. Aside from being falsified by the actual data, Flyvbjerg points out the non-monetary opportunity costs through the example of the Sydney Opera House. It's architect, Dane Jorn Utzon, won the Pritzker Prize (the Nobel of architecture) in 2003 for the Sydney Opera House. It is his only major work - the catastrophic delays and cost overruns destroyed his career. Contrast with Frank Gehry, another inspired architect, and it looks like management's bungling of the Opera House probably cost us half a dozen gorgeous landmarks.

Survival of the Unfittest

The prevailing attitude that it is perfectly acceptable to lie about and then badly manage megaprojects leads to a weird scenario where worse projects are more likely to be chosen. Consider two competing projects, one with honest and competent management, and one with dishonest and incompetent management. The costs look lower for the latter project, and the benefits look higher, and the choosers between them probably expect them to both be over budget and behind schedule by about the same amount. Therefore we systematically make worse decisions about what projects to build.

Light at the End of the Tunnel

Fortunately there are bright spots. During the Obama administration these failings were identified as an important policy area for the US government. It is now much more common for a megaproject failure to result in consequences for leadership, like the CEOs of BP and Deepwater Horizon, or Airbus and the A380 superjumbo. There are megaprojects that go well and serve as examples of how to do it right, like the Guggenheim Museum in Bilbao. Lastly, there is scattered adoption of varying levels of good practices, like reference class forecasting and independent forecasting.

End

For those interested in learning more, Oxford has Major Programme Management at the masters level. In books there is the Oxford Handbook of Megaproject Management, and Megaproject Planning and Management: Essential Readings, both from Flyvbjerg. I have read neither, and both are collections of papers - I may just hunt them down independently, for budgetary reasons.

Less Competition, More Meritocracy?

Analysis of the paper: Less Competition, More Meritocracy (hat tip: Marginal Revolution: Can Less Competition Mean More Meritocracy?)

Epistemic Status: Consider the horse as if it was *not* a three meter sphere

Economic papers that use math to prove things can point to interesting potential results and reasons to question one's intuitions. What is frustrating is the failure to think outside of those models and proofs, analyzing the practical implications.

In this particular paper, the central idea is that when risk is unlimited and free, ratcheting up competition dramatically increases risk taken. This introduces sufficient noise that adding more competitors can make the average winner less skilled. At the margin, adding additional similar competitors to a very large pool has zero impact. Adding competitors with less expected promise makes things worse.

This can apply in the real world. The paper provides a good example of a very good insight that is then proven 'too much,' and which does not then question or vary its assumptions in the ways I would find most interesting.

I. The Basic Model and its Central Point

Presume some number of job openings. There are weak candidates and strong candidates. Each candidate knows if they are strong or weak, but not how many other candidates are strong, nor do those running the contest know how many are strong.

The goal of the competition is to select as many strong candidates as possible. Or formally, to maximize [number of strong selected – number of weak selected], which is the same thing if the number of candidates is fixed, but is importantly different later when the number of selected candidates can vary. Each candidate performs and is given a score, and for an N-slot competition, the highest N scores are picked.

By default, strong candidates score X and weak candidates score Y, X>Y, but each candidate can also take on as much risk as they wish, with any desired distribution of scores, so long as their score never goes below zero.

The paper then does assumes reflexive equilibrium, does math and proves a bunch of things that happen next. The math checks out; I duplicated the results intuitively.

There are two types of equilibrium.

In the first type, *concession equilibria,* strong candidates take no risk and are almost always chosen. Weak candidates take risk to try and beat other weak candidates, but attempting to beat strong candidates isn't worthwhile. This allows strong candidates to take zero risk.

In the second type, *challenge equilibria,* weak candidates attempt to be chosen over strong candidates, forcing strong candidates to take risk.

If I am a weak candidate, I can be at least (Y/X) as likely as a strong candidate to be selected by copying their strategy with probability (Y/X) and scoring 0 otherwise. This seems close to optimal in a challenge equilibria.

Adding more candidates, strong or weak, risks shifting from a concession to a challenge equilibria. Each additional candidate, of any strength, makes challenge a better option relative to concession.

If competition is ‘insufficiently intense’ then we get a concession equilibria. We successfully identify every strong candidate, at the cost of accepting some weak ones. If competition is ‘too intense’ we lose that. The extra candidate that tips us over the edge makes things much worse. After that, quantity does not matter, only the ratio of weak candidates to strong.

Even if search is free, and you continue to sample from the same pool, hitting the threshold hurts you, and further expansion does nothing. Interviewing one million people for ten jobs, a tenth of which are strong, is not better than ten thousand, or even one hundred. Ninety might be better.

Since costs are never zero (and rarely negative), and the pool usually degrades as it expands, this argues strongly for limited competitions with weaker selection criteria, including via various hacks to the system.

II. What To Do, and What This Implies, If This Holds

The paper does a good job analyzing what happens if its conditions hold.

If one has a fixed set of positions to fill (winners to pick) and wants to pick the maximum number of strong candidates, with no cost to expanding the pool of candidates, the ideal case is to pick the maximum number of strong candidates that maintains a concession equilibrium. With no control (by assumption) over who you select or how to select them, this is the same as picking the maximum number of candidates that maintains a concession equilibrium, no matter what decrease in quality you might get while expanding the pool.

The tipping point makes this a Price Is Right style situation. Get as close to the number as possible without going over. Going over is quite bad, worse than a substantial undershoot.

One can think of probably not interviewing enough strong candidates, and probably hiring some weak candidates, as the price you must pay to be allowed to sort strong candidates from weak candidates – **you need to ‘pay off’ the weak ones to not try and fool the system**. An extra benefit is that even as you fill all the slots, *you know who is who,* which can be valuable information in the future. Even if you’re stuck with them, better to know that.

A similar dynamic comes if choosing how many candidates to select from a fixed pool, or when choosing both candidate and pool sizes.

If one attempts to only have slots for strong candidates, under unlimited free risk taking, you guarantee a challenge equilibria. Your best bet will therefore probably be to pick enough candidates from the pool to create a concession equilibrium, just like choosing a smaller candidate pool.

The paper considers hiring a weak candidate as a -1, and hiring a strong candidate as a +1. The conclusions don’t vary much if this changes, since there are lots of other numerical knobs left unspecified that can cancel this out. But it is worth noting that in most cases the ratio is far less favorable than that. The default is that one good hire is

far less good than one bad hire is bad. True bad hires are rather terrible (as opposed to all right but less than the best).

Thus, when the paper points out that it is sometimes impossible to reliably break 50% strong candidates under realistic conditions, no matter how many people are interviewed and how many slots are given out, they *underestimate* the chance that the system breaks down entirely into no contest at all, and no production.

What is the best we can do, if all assumptions hold?

The minimum portion of weak candidates accepted scales linearly with their presence in the pool, and with how strongly they perform relative to strong candidates. Thus we set the pool size such that this fills out the pool with some margin of error.

That is best if we set the pool size but nothing else. The paper considers college admissions. A college is advised to solve for which candidates are above a fixed threshold, then choose at random from those above the threshold (which is a suggestion one would only make in a paper with zero search costs, since once you have enough worthy candidates *you can stop searching,* but shrug.) Thus, we can always choose to arbitrarily limit the pool.

In practice, attempting this would change the pool of applicants. In a way you won't like. You are more attractive to weak candidates and less attractive to strong ones. Weak candidates flood in to 'take their shot,' causing a vicious cycle of reputation and pool decay. You've not a good reach school or a safe school for a strong candidate, so why bother? If other colleges copy you, students respond by investing less in becoming strong and more in *sending out all the applications,* and the remaining strong candidates remain at risk.

True reflexive equilibria almost never exist, given the possible angles of response, and differences between people's knowledge, preferences and cognition.

III. Relax Reflective Equilibrium

Even if it is common knowledge that only two candidate strengths exist, and all candidates of each type are identical (which they aren't), they will get different information and react differently, destroying reflexive equilibrium.

Players will not expect all others to jump with certainty between equilibria at some size threshold. Because they won't. Which creates different equilibria.

Some players don't know game theory, or don't pay attention to strategy. Those players, as a group, lose. Smart game theory always has the edge.

An intuition pump: Learning game theory is costly, so the equilibrium requires it to pay off. Compare to the efficient market hypothesis.

Some weak candidates will always attempt to pass as strong candidates. There is a gradual shift from most not doing so to almost everyone doing so. More weak candidates steadily take on more risk. Eventually most of them mostly take on large risk to do their impression of a strong candidate. Strong candidates slowly start taking more risk more often as they sense their position becoming unsafe.

Zero risk isn't stable anyway without continuous skill levels. Strong candidates notice that *exactly zero* risk puts them behind candidates who take on extra tail risk to get

epsilon above them. Zero risk is a default strategy, so beating that baseline is wise.

Now those doing this try to outbid each other, until strong candidates lose to weak candidates at least sometimes. This risk will cap out very low if strong candidates consider the risk of losing at around their average performance to also be minuscule, but it will have to exist. Otherwise, there's an almost free action in making one's poor performances worse, since they are already losing to almost all other strong candidates, and doing that allows one to make their stronger performances better and/or more likely.

The generalization of this rule is that **whenever you introduce a possible outcome into the system, and provide *any* net benefit to *anyone* if they do things that make the outcome more likely, there is now a chance that the outcome happens.** Even if the outcome is 'divorce,' 'government default,' 'forced liquidation,' 'we both drive off the cliff' or 'nuclear war.' It probably also isn't epsilon. While risk is near epsilon, taking actions that increase risk will look essentially free, so *until the risk is big enough to matter* it will keep increasing. Therefore, every risk isn't only *possible.* Every risk *will matter.* Given enough time, someone will miscalculate, and Murphy's Law ensues.

Future post: Possible bad outcomes are really bad.

Stepping back, the right strategy for each competitor will be to guess the performance levels that efficiently translate into wins, making sure to maximally bypass levels others are likely to naively select (such as zero risk strategies), and generally play like they're in a variation of the game of Blotto.

A lot of these results are driven by discrete skill levels, so let's get rid of those next.

IV. Allow Continuous Skill Levels

Suppose instead of two skill levels, each player has their own skill level, and a rough and noisy idea where they lie in the distribution.

Each player has resources to distribute across probability. Success is increasing as a function of performance. Thinking players aim for performance levels they believe are efficient, and do not waste resources on performance levels that matter less.

All competitors also know that the chance of winning with low performance is almost zero. The value of additional performance probably gradually increases (positive second derivative) until it peaks at an inflection point, and then starts to decline as success starts to approach probability one. There may be additional quirky places in the distribution where extra performance is especially valuable. This exact curve won't be known to anyone, different players will have different guesses partly based on their own abilities, and ability levels are continuous.

A sufficiently strong candidate, who expects their average performance to be above the inflection point, should take no risk. A weaker candidate should approximate the inflection point, and risk otherwise scoring a zero performance to reach that point. Simple.

If the distribution of skill levels is bumpy, what happens then? We have strong candidates and weak candidates (e.g. let's say college graduates and high school graduates, or some have worked in the field and some haven't, or whatever) so

there’s a two-peak distribution of skill levels. Unless people are badly misinformed, we’ll still get a normal-looking distribution. If the two groups calculate very different expected thresholds, we’ll see two peaks.

In general, but not always, enough players will miscalculate or compete for the ‘everyone failed’ condition that trying to do so is a losing play. Occasionally there will be good odds to hoping enough others aim too high and miss.

Rather than have a challenge and a concession equilibrium, we have a threshold equilibrium. Everyone has a noisy estimate of the threshold they need. Those capable of reliably hitting the threshold take no risk, and usually make it. Those not capable of reliably hitting the threshold risk everything to make the threshold as often as possible.

Note that this equilibrium holds, although it may contain no one above the final threshold. If everyone aims for what they think is good-enough performance, aiming for less is almost worthless, and aiming for much more is mostly pointless, and threshold adjusts so that the expected number of threshold performances is very close to the number of slots.

More competition raises the threshold, forcing competitors to take on more risk, until everyone is using the same threshold strategy and success is purely proportional to skill. Thus, in a large pool, we once again have expanding the pool as a bad idea if it weakens average skill, even if search and participation costs for all are free.

In a small pool, the strongest candidates are ‘wasting’ some of their skill on less efficient outcomes beyond their best estimate of the threshold.

This ends up being similar to the challenge case, except that there is no inflection point where things suddenly get worse. You never expect to lose from expanding the pool while maintaining quality. Instead, things slowly get better as you waste less work at the top of the curve, so the value of adding more similar candidates quickly approaches zero.

The new intuition is, given low enough search costs, we should add equally strong potential candidates until we are confident everyone is taking risk, rather than stopping just short of causing stronger candidates to take risk. If participation is costly to you and/or the candidates, you should likely stop short of that point.

The key intuitive question to ask is, if a candidate was the type of person you want, would they be so far ahead of the game as to be obviously better than the current expected marginal winner? Would they be able to crush a much bigger pool, and thus be effectively wasting lots of effort? If and only if that’s true, there’s probably benefit to expanding your search, so you get more such people, and it’s a question of whether it is worth the cost.

The other strong intuition is that once your marginal applicant pool is lower in average quality than your average pool, that will always be a high cost, so focus on quality over quantity.

This suggests another course of action…

V. Multi-Stage Process

Our model tells us that average quality of winners is, given a large pool, a function of the *average* quality of our base pool.

But we have a huge advantage: This whole process *is free.*

Given that, it seems like we should be able to be a bit more clever and complex, and do better.

We can improve if we can get a pool of candidates that has a higher *average quality* than our original candidate pool, but which is large enough to get us into a similar equilibrium. Each candidate's success is proportional to their skill level, so our average outcome improves.

We already have a selection process that does this. We know our winners will be on average better than our candidates. So why not use that to our advantage?

Suppose we did a multi-stage competition. Before, we would have had 10 applicants for 1 slot. Expanding that to 100 applicants won't do us any good directly, because of risk taking. But running 10 competitions with 10 people each, then pitting those 10 winners against each other, will improve things for us.

By using this tactic multiple times, we can do quite a bit better. Weaker candidates will almost never survive multiple rounds.

What happened here?

We cheated. We forced candidates to take observable, *uncorrelated* risks in each different round. We destroyed the rule that risk taking is free and easy, and assumed that a lucky result in round 1 won't help you in round 2.

If a low-skill person can *permanently* mimic in all ways a high-skill person, and we observe that success, *they are high skill now!* A worthy winner. If they can't, then they fall back down to Earth on further observation. This should make clear why the idea of *unlimited cheap and exactly controlled risk* is profoundly bizarre. A test that works that way is a rather strange test.

So is a test that costs nothing to administer. You get what you pay for.

The risk is that risk-taking takes the form of 'guess the right approach to the testing process' and thus test scores are correlated without having to link back to skill.

This is definitely a thing.

During one all-day job interview, I made several fundamental interview-skill mistakes that hurt me in multiple sessions. If I had fixed those mistakes, I would have done much better all day, *but would not have been much more skilled at what they were testing for.* A more rigorous or multi-step process could have only done so much. To get better information, they would have had to add a *different kind of test.* That would risk introducing bad noise.

This seems typical of similar contests and testing methods designed to find strong candidates.

A more realistic model would introduce costs to participation in the search process, for all parties. You'd have another trade-off between having noise be correlated versus

minimizing its size, making more rounds of analysis progressively less useful.

Adding more candidates to the pool now clearly is good at first and then turns increasingly negative.

VI. Pricing People Out

There are two realistic complications that can help us a lot.

The first is pricing people out. Entering a contest is rarely free. I have been fortunate that my last two job interviews were at Valve Software and Jane Street Capital. Both were exceptional companies looking for exceptional people, and I came away from both interviews feeling like I'd had a very fun and very educational experience, in addition to leveling up my interview skills. So *those particular* interviews felt free or better. But most are not.

Most are more like when I applied to colleges. Each additional college meant a bunch of extra work plus an application fee. Harvard does not want to admit a weak candidate. If we ignore the motivation to show that you have lots of applications, Harvard would prefer that weak candidates not apply. It wastes time, and there's a non-zero chance one will gain admission by accident. If Harvard taxes applications, by requiring additional effort or raising the fee, they will drive weak applicants away and strengthen their pool, improving the final selections.

Harvard also does this by making Harvard hard. A sufficiently weak candidate should not *want* to go to Harvard, because they will predictably flunk out. Making Harvard harder, the way MIT is hard, would make their pool higher quality once word got out.

We can think of some forms of hazing, or other bad experiences for winners of competitions, partly as a way to discourage weak candidates from applying, and also partly as an additional test to drive them out.

Ideally we also *reduce risk taken.*

A candidate has uncertainly in how strong they are, and how much they would benefit from the prize. If being a stronger candidate is correlated with benefiting from winning, a correct strategy becomes to take less or no risk. If taking a big risk causes me to win when I would otherwise lose, *I won a prize I don't want.* If taking a big risk causes me to lose, *I lost a prize I did want.* That pushes me heavily towards lowering my willingness to take risk, which in turn lowers the competition level and encourages me to take less risk still. Excellent.

VII. Taking Extra Risk is Hard

Avoiding risk is also hard.

In the real world, there is a 'natural' amount of risk in any activity. One is continuously offered options with varying risk levels.

Some of these choices are big, some small. Sometimes the risky play is 'better' in an expected value sense, sometimes worse.

True max-min strategies that avoid even minimal risks decline even small risks that would cancel out over time. This is expensive.

If one wants to maximize risk at all costs, one ends up doing the more risky thing every time and takes bad gambles. This is also expensive.

It is a hard problem to get the best outcome given one's desired level of risk, or to maximize the chance of exceeding some performance threshold, even with no opponent. In games with an opponent who wants to beat you and thus has the opposite incentives of yours (think football) it gets harder still. Real world performances are notoriously terrible.

There are two basic types of situations with respect to risk.

Type one is where adding risk is expensive. There is a natural best route to work or line of play. There are other strategies that overall are worse, but have bigger upside, such as taking on particular downside tail risks in exchange for tiny payoffs, or hoping for a lucky result. In the driving example, one might take an on average slower route that has variable amounts of traffic, or one might drive faster and risk an accident or speeding ticket.

Available risk is limited. If I am two hours away by car, I might be able to do something reckless and maybe get there in an hour and forty five minutes, but if I have to get there in an hour, it's not going to happen.

I can hope to ever overcome only a limited skill barrier. If we are racing in the Indianapolis 500, I might try to win the race by skipping a pit stop, or passing more aggressively to make up ground, or choosing a car that is slightly faster but has more engine trouble. But if my car combined with my driving skill is substantially slower than yours (where substantially means a minute over several hours) and your car doesn't crash or die, *I will never beat you.*

If I had taken the math Olympiad exam (the USAMO) another hundred times, I might have gotten a *non-zero* score sometimes, but I was never getting onto the team. Period.

In these situations, reducing risk beyond the 'natural' level may not even be possible. If it is, it will be increasingly expensive.

Type two is where giant risks are the default, then sacrifices are made to contain those risks. Gamblers who do not pay attention to risk will always go broke. To be a winning gambler, one can either be lucky and retain large risk, or one can be skilled and pay a lot of attention to containing risk. In the long term, containing risk, including containing risk by ceasing to play at all, is the only option.

Competitors in type two situations must be evaluated explicitly on their risk management, or on very long term results, or any evaluation is worthless. If you are testing for good gamblers and only have one day, you pay some attention to results but more attention to the logic behind choices and sizing. Tests that do otherwise get essentially random results, and follow the pattern where reducing the applicant pool improves the quality of the winners.

Another note is that the risks competitors take can be *correlated across competitors* in many situations. If you need a sufficiently high rank rather than a high raw score, those who take risks should seek to take *uncorrelated* risks. Thus, in stock market or gambling competitions, the primary skill often is in doing something no one else would think to do, rather than in picking a high expected value choice. Sometimes that's what real risk means.

VIII. Central Responses

There are also four additional responses by those running the competition, that are worth considering.

The first response is to observe a competitor's level of risk taking and test optimization, and penalize too much (or too little). This is often quite easy. Everyone knows what a safe answer to 'what is your greatest weakness' looks like, bet size in simulations is transparent, and so on. If you respond to things going badly early on with taking a lot of risk, rather than being responsible, will you do that with the company's money?

A good admissions officer at a college mostly knows instantly which essays had professional help and which resumes are based on statistical analysis, versus who lived their best life and then applied to college.

A good competition design gives you the opportunity to measure these considerations.

Such contests should be anti-inductive, if done right, with the really sneaky players playing on higher meta levels. Like everything else.

The second response is to vary the number of winners based on how well competitors do. This is the default.

If I interview three job applicants and all of them show up hung over, I need to be pretty desperate to take the one who was *less* hung over, rather than call in more candidates tomorrow. If I find three great candidates for one job, I'll do my best to find ways to hire all three.

Another variation is that I have an insider I know well as the default winner, and the application process is to see if I can do better than that, and to keep the insider and the company honest, so again it's mostly about crossing a bar.

The third response is that often there isn't even a 'batch' of applications. There is only a series of permanent yes/no decisions until the position is filled. This is the classic problem of finding a spouse or a secretary, where you can't easily go back once you reject someone. Once you have a sense of the distribution of options, you're effectively looking for 'good enough' at every step, and that requirement doesn't move much until time starts running out.

Thus, most contests *that care mostly about finding a worthy winner* are closer to threshold requirements than they look. This makes it very difficult to create a concession equilibrium. If you show up and aren't good enough to beat continuing to search, your chances are very, very bad. If you show up and are *are* good enough to beat continuing to search, your chances are very good. The right strategy becomes either to aim at this threshold, or if the field is large you might need to aim higher. You can never keep the field small enough to keep the low-skill players honest.

The fourth response is to punish sufficiently poor performance. This can be as mild as in-the-moment social embarrassment – Simon mocking aspirants in American Idol. It can be as serious as 'you're fired,' either from the same company (you revealed you're not good enough for your current job, or your upside is limited), or from another company (how dare you try to jump ship!). In fiction a failed application can

be lethal. Even mild retaliation is very effective in improving average quality (and limiting the size) of the talent pool.

IX. Practical Conclusions

We don't purely want the best person for the job. We want a selection process that balances search costs, for all concerned, with finding the best person and perhaps getting your applicants to improve their skill.

A weaker version of the paper's core take-away heuristic seems to hold up under more analysis: ***There is a limit to how far expanding a search helps you at all, even before costs.***

Rule 1: Pool quality on the margin usually matters more than quantity.

Bad applicants that can make it through are more bad than they appear. Expanding the pool's quantity at the expense of average quality, once your supply of candidates isn't woefully inadequate, is usually a bad move.

Rule 2: Once your application pool probably includes enough identifiable top-quality candidates to fill all your slots, up to your ability to differentiate, stop looking.

A larger pool will make your search more expensive and difficult for both you and them, add more regret because choices are bad, and won't make you more likely to choose wisely.

Note that this is a later stopping point than the paper recommends. The paper says you should stop *before* you fill all your slots, such that weak applicants are encouraged not to represent themselves as strong candidates.

Also note that this rule has two additional requirements. It requires the good candidates be identifiable, since if some of them will blow it or you'll blow noticing them, that doesn't help you. It also requires that there not be outliers waiting to be discovered, that you would recognize if you saw them.

Another, similar heuristic that is also good is, **make the competition just intense enough that worthy candidates are worried they won't get the job. Then stop.**

Rule 3: Weak candidates must either be driven away, or rewarded for revealing themselves. If weak candidates can successfully fake being strong, it is worth a lot to ensure that this strategy is punished.

Good punishments include application fees, giving up other opportunities or jobs, long or stressful competitions, and punishments for failure ranging from mild in-the-room social disapproval or being made to feel dumb, up to major retaliation.

Another great punishment is to give less rewards to success if it is by a low skilled person. If their prize is something they can't use – they'll flunk out, or get fired quickly, or similar – then they will be less inclined to apply.

Reward for participation is probability of success times reward for success, while cost is mostly fixed. Tilt this enough and your bad-applicant problem clears up.

Fail to tilt this enough, and you have a big lemon problem on multiple levels. Weak competitors will choose your competition over others, giving strong applicants less reason to bother both in terms of chance of winning, and desire to win. Who wants to win only to be among a bunch of fakers who got lucky? That's no fun and it's no good for your reputation either.

It will be difficult to punish weak candidates *for faking being strong* versus punishing them in general. But if you can do it, that's great.

The flip side is that we can reward them for being honest. That will often be easier.

Preventing a rebellion of the less skilled is a constraint on mechanism design. We must either appease them, or wipe them out.

Rule 4: Sufficiently hard, high stakes competitions that are vulnerable to gaming and/or resource investment are highly toxic resource monsters.

This is getting away from the paper's points, since the paper doesn't deal with resource costs to participation or search, but it seems quite important.

In some cases, we *want* these highly toxic resource monsters. We like that every member of area sports team puts the rest of their life mostly on hold and focuses on winning sporting events. The test is exactly what we want them to excel at. We also get to use the trick of testing them in discrete steps, via different games and portions of games, to prevent 'risk' from playing too much of a factor.

In most cases, where the match between test preparation, successful test strategies and desired skills is not so good, this highly toxic resource monster is very, very bad.

Consider school, or more generally childhood. The more we reward good performance on a test, and punish failure, the more resources are eaten alive by the test. In the extreme, all of most child's experiences and resources, and even those of their parents, become eaten. From discussions I've had, much of high school in China has something remarkably close to this, as everything is dropped *for years* to cram for a life-changing college entrance exam.

Rule 5: Rewards must be able to step outside of a strict scoring mechanism.

Any scoring mechanism is vulnerable to gaming and to risk taking, and to Goodhart's Law. To avoid everyone's motivation, potentially their entire life and being, being subverted, we need to be rewarding and punishing from the outside looking in on what is happening. This has to carry enough weight to be competitive with the prizes themselves.

Consider this metaphor.

If the real value of many journeys is the friends you made along the way, that can be true in both directions. Often one's friends, experiences and lessons end up dwarfing in importance the prize or motivation one started out with; frequently we need a McGuffin and restrictions that breed creativity and focus to allow coordination, more than any prize.

It also works *the other way.* The value of your friends can be that they motivate and help you to be worthy of friendship, to do and accomplish things. **The reason we took the journey the right way was so that we would make friends along**

it. This prevents us from falling to Goodhart's Law. We don't narrow in on checking off a box. Even in a pure competition, like a Magic tournament, we know the style points matter, and we know that it matters whether we think the style points matter, and so on.

The existence of the social, of various levels and layers, *the ability to step outside the game,* and *the worry about unknown unknowns,* is what guards systems from breakdown under the pressure of metrics. Given any utility function we know about, however well designed, and sufficient optimization pressure, things end badly. You need to preserve the value of unknown unknowns.

This leads us to:

Rule 6: Too much knowledge by potential competitors can be very bad.

The more competitors do the 'natural' thing, that maximizes their expected output, the better off we usually are. The less they know about how they are being evaluated, on what levels, with what threshold of success, the less they can game the system, and the less success depends on gaming skill or luck.

All the truly perverse outcomes came from scenarios where competitors knew they were desperadoes, and taking huge risks was not actually risky for them.

Having a high threshold is only bad *if competitors know about it.* If they don't know, it can't hurt you. If they *suspect* a high threshold, *but they don't know,* that mitigates a lot of the damage. In many cases, the competitor is better served by playing to succeed in the worlds where the threshold is low, and accept losing when the threshold is unexpectedly high, which means doing exactly what you want. More uncertainty also makes the choices of others less certain, which makes situations harder to game effectively.

Power hides information. Power does not reveal its intentions. This is known, and the dynamics explored here are part of *why.* You want people optimizing for things you won't even be aware of, or don't care about, but which *they think you might be aware of and care about.* You want to avoid them trying too hard to game the things you *do* look at, which would also be bad. You make those in your power worry at every step that if they try anything, or fail in any way, it could be what costs them. You cause people to want to curry favor. You also allow yourself to alter the results, if they're about to come out 'wrong'. The more you reveal about how you work, the less power you have. In this case, the power to find worthy winners.

This is in addition to the fact that some considerations that matter are not legally allowed to be considered, and that lawsuits might fly, and other reasons why decision makers ensure that no one knows what they were thinking.

Thus we must work even harder to reward those who explain themselves and thereby help others, and who realize that the key hard thing is, as Hagbard Celine reminds us, to *avoid* power.

But still get things done.

From Personal to Prison Gangs: Enforcing Prosocial Behavior

This post originally appeared here; I've updated it slightly and posted it here as a follow-up to this post.

David Friedman has a fascinating book on alternative legal systems. One chapter focuses on prison law - not the nominal rules, but the rules enforced by prisoners themselves.

The unofficial legal system of California prisoners is particularly interesting because it underwent a phase change sometime after the 1960's.

Prior to the 1960's, prisoners ran on a decentralized code of conduct - various unwritten rules roughly amounting to "mind your own business and don't cheat anyone". Prisoners who kept to the code were afforded some respect by their fellow inmates. Prisoners who violated the code were ostracized, making them fair game for the more predatory inmates. There was no formal enforcement; the code was essentially a reputation system.

Sometime after the 1960's, that changed. During the code era, California's total prison population was only about 5000, with about 1000 inmates in a typical prison. That's quite a bit more than Dunbar's number, but still low enough for a reputation system to work through second-order connections. By 1970, California's prison population had ballooned past 25000; today it is over 170000. The number of prisons also grew, but not nearly as quickly as the population, and today's prisoners frequently move across prisons anyway. In short, a decentralized reputation system became untenable. There were too many other inmates to keep track of.

As the reputation system collapsed, a new legal institution grew to fill the void: prison gangs. Under the gang system, each inmate is expected to affiliate with a gang (though most are not formal gang members). The gang will explain the rules, often in written form, and enforce them on their own affiliates. When conflict arises between affiliates of different gangs, the gang leaders negotiate settlement, with gang leaders enforcing punishments on their own affiliates. (Gang leaders are strongly motivated to avoid gang-level conflicts.) Rather than needing to track reputation of everyone individually, inmates need only pay attention to gangs at a group level.
Of course, inmates need some way to tell who is affiliated with each gang - thus the rise of racial segregation in prison. During the code era, prisoners tended to associate by race and culture, but there was no overt racial hostility and no hard rules against associating across race. But today's prison gangs are highly racially segregated, making it easy to recognize the gang affiliation of individual inmates. They claim territory in prisons - showers or ball courts - and enforce their claims, resulting in hard racial segregation.

The change from a small, low-connection prison population to a large, high-connection population was the root cause. That change drove a transition from a decentralized, reputation-based system to prison gangs. This, in turn, involved two further transitions. First, a transition from decentralized, informal unwritten rules to formal written rules with centralized enforcement. Second, a transition from individual to group-level identity, in this case manifesting as racial segregation.

Generalization

This is hardly unique to prisons. The pattern is universal among human institutions. In small groups, everybody knows everybody. Rules are informal, identity is individual. But as groups grow:

- Rules become formal, written, and centrally enforced
- Identity becomes group-based.

Consider companies. I work at a ten-person company. Everyone in the office knows everyone else by name, and everyone has some idea of what everyone else is working on. We have nominal job titles, but everybody works on whatever needs doing. Our performance review process is to occasionally raise the topic in weekly one-on-one meetings.

Go to a thousand or ten thousand person company, and job titles play a much stronger role in who does what. People don't know everyone, so they identify others by department or role. They understand what a developer or a manager does, rather than understanding what John or Allan does. Identity becomes group-based. At the same time, hierarchy and bureaucracy are formalized.

The key parameter here is number of interactions between each pair of people (you should click that link, it's really cool). In small groups, each pair of people has many interactions, so people get to know each other. In large groups, there are many one-off interactions between strangers. Without past interactions to fall back on, people need other ways to figure out how to interact with each other. One solution is formal rules, which give guidance on interactions with anyone. Another solution is group-based identity - if I know how to interact with lawyers at work in general, then I don't need to know each individual lawyer.

In this regard, prisons and companies are just microcosms of society in general.

Society

At some point over the past couple hundred years, society underwent a transition similar to that of the California prison system.

In 1800, people were mostly farmers, living in small towns. The local population was within an order of magnitude of Dunbar's number, and generally small enough to rely on reputation for day-to-day dealings.

Today, that is not the case [citation needed].

Just as in prisons and companies, we should expect this change to drive two kinds of transitions:

- A transition from informal, decentralized rules to formal, written, centrally-enforced rules.
- A transition from individual to group-level identity.

This can explain an awful lot of the ways in which society has changed over the past couple hundred years, as well as how specific social institutions evolve over time. To take just a few examples…

- Regulation. As people have more one-off interactions, reputation becomes less tenable, and we should expect formal regulation to grow. Conversely, regulations are routinely ignored among people who know each other.
- Litigation. Again, with more one-off interactions, we should expect people to rely more on formal litigation and less on informal settlement. Conversely, people who interact frequently rarely sue each other - and when they do, it's expected to mess up the relationship.
- Professional licensing. Without reputation, people need some way to signal that they are safe to hire. We should expect licensing to increase as pairwise interactions decrease.
- Credentialism. This is just a generalization of licensing. As reputation fails, we should expect people to rely more heavily on formal credentials - "you are your degree" and so forth.
- Stereotyping. Without past interactions with a particular person, we should expect people to generalize based on superficially "similar" people. This could be anything from the usual culprits (race, ethnicity, age) to job roles (actuaries, lawyers) to consumption signals (iphone, converse, fancy suit).
- Tribalism. From nationalism to sports fans to identity politics, an increasing prevalence of group-level identity means an increasing prevalence of tribal behavior. In particular, I'd expect that social media outlets with more one-off or low-count interactions are characterized by more extreme tribalism.
- Standards for impersonal interactions. "Professionalism" at work is a good example.

I've focused mostly on negative examples here, but it's not all bad - even some of these examples have upsides. When California's prisons moved from an informal code to prison gangs, the homicide rate dropped like a rock; the gangs hate prison lockdowns, so they go to great lengths to prevent homicides. Of course, gangs have lots of downsides too. The point which generalizes is this: bodies with centralized power have their own incentives, and outcomes will be "good" to exactly the extent that the incentives of the centralized power align with everybody else' incentives and desires.

Consider credentialism, for example. It's not all bad - to the extent that we now hire based on degree rather than nepotism, it's probably a step up. But on the other hand, colleges themselves have less than ideal incentives. Even setting aside colleges' incentives, the whole credential system shoehorns people into one-size-fits-all solutions; a brilliant patent clerk would have a much more difficult time making a name in physics today than a hundred years ago.

Takeaway

Of course, all of these examples share one critical positive feature: they scale. That's the whole reason things changed in the first place - we needed systems which could scale up beyond personal relationships and reputation.

This brings us to the takeaway: what should you do if you want to change these things? Perhaps you want a society with less credentialism, regulation, stereotyping, tribalism, etc. Maybe you like some of these things but not others. Regardless, surely there's something somewhere on that list you're less than happy about.
The first takeaway is that these are not primarily political issues. The changes were driven by technology and economics, which created a broader social graph with fewer repeated interactions. Political action is unlikely to reverse any of these changes; the

equilibrium has shifted, and any policy change would be fighting gravity. Even if employers were outlawed from making hiring decisions based on college degree, they'd find some work-around which amounted to the same thing. Even if the entire federal register disappeared overnight, de-facto industry regulatory bodies would pop up. And so forth.

So if we want to e.g. reduce regulation, we should first focus on the underlying socioeconomic problem: fewer interactions. A world of Amazon and Walmart, where every consumer faces decisions between a million different products, is inevitably a world where consumers do not know producers very well. There's just too many products and companies to keep track of the reputation of each. To reduce regulation, first focus on solving that problem, scalably. Think amazon reviews - it's an imperfect system, but it's far more flexible and efficient than formal regulation, and it scales.

Now for the real problem: online reviews are literally the only example I could come up with where technology offers a way to scale-up reputation-based systems, and maybe someday roll back centralized control structures or group identities. How can we solve these sorts of problems more generally? Please comment if you have ideas.

Reframing Superintelligence: Comprehensive AI Services as General Intelligence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf?asd=sa

Since the CAIS technical report is a gargantuan 210 page document, I figured I'd write a post to summarize it. I have focused on the earlier chapters, because I found those to be more important for understanding the core model. Later chapters speculate about more concrete details of how AI might develop, as well as the implications of the CAIS model on strategy. ETA: This comment provides updates based on more discussion with Eric.

The Model

The core idea is to look at the pathway by which we will develop general intelligence, rather than assuming that at some point we will get a superintelligent AGI agent. To predict how AI will progress in the future, we can look at how AI progresses currently -- through research and development (R&D) processes. AI researchers consider a problem, define a search space, formulate an objective, and use an optimization technique in order to obtain an AI system, called a *service*, that performs the task.

A service is an AI system that delivers bounded results for some task using bounded resources in bounded time. Superintelligent language translation would count as a service, even though it requires a very detailed understanding of the world, including engineering, history, science, etc. Episodic RL agents also count as services.

While each of the AI R&D subtasks is currently performed by a human, as AI progresses we should expect that we will automate these tasks as well. At that point, we will have automated R&D, leading to recursive *technological* improvement. This is *not* recursive *self*-improvement, because the improvement comes from R&D services creating improvements in basic AI building blocks, and those improvements feed back into the R&D services. All of this should happen before we get any powerful AGI agents that can do arbitrary general reasoning.

Why Comprehensive?

Since services are focused on particular tasks, you might think that they aren't general intelligence, since there would be some tasks for which there is no service. However, pretty much everything we do can be thought of as a task -- including the task of creating a new service. When we have a new task that we would like automated, our service-creating-service can create a new service for that task, perhaps by training a new AI system, or by taking a bunch of existing services and putting them together, etc. In this way, the collection of services can perform any task, and so as an aggregate is generally intelligent. As a result, we can call this

Comprehensive AI Services, or CAIS. The "Comprehensive" in CAIS is the analog of the "General" in AGI. So, we'll have the capabilities of an AGI agent, before we can actually make a monolithic AGI agent.

Isn't this just as dangerous as AGI?

You might argue that each individual service must be dangerous, since it is superintelligent at its particular task. However, since the service is optimizing for some *bounded* task, it is not going to run a long-term planning process, and so it will not have any of the standard convergent instrumental subgoals (unless the subgoals are helpful for the task before reaching the bound).

In addition, all of the optimization pressure on the service is pushing it towards a particular narrow task. This sort of strong optimization tends to *focus* behavior. Any long term planning processes that consider weird plans for achieving goals (similar to "break out of the box") will typically not find any such plan and will be eliminated in favor of cognition that will actually help achieve the task. Think of how a racecar is optimized for speed, while a bus is optimized for carrying passengers, rather than having a "generally capable vehicle".

It's also worth noting what we mean by superintelligent here. In this case, we mean that the service is extremely *competent* at its assigned task. It need not be *learning* at all. We see this distinction with RL agents -- when they are trained using something like PPO, they are learning, but at test time you can simply execute them without any PPO and they will perform the behavior they previously learned and won't change that behavior at all.

(My opinion: I think this isn't engaging with the worry with RL agents -- typically, we're worried about the setting where the RL agent is learning or planning at test time, which can happen in learn-to-learn and online learning settings, or even with vanilla RL if the learned policy has access to external memory and can implement a planning process separately from the training procedure.)

On a different note, you might argue that if we analyze the system of services as a whole, then it certainly looks generally intelligent, and so should be regarded as an AGI agent. However, "AGI agent" usually carries the anthropomorphic connotation of VNM rationality / expected utility maximization / goal-directedness. While it seems possible and even likely that each individual service can be well-modeled as VNM rational (albeit with a bounded utility function), it is *not* the case that a system of VNM rational agents will itself look VNM rational -- in fact, game theory is all about how systems of rational agents have weird behavior.

In addition, there are several aspects of CAIS that make it more safe than a classic monolithic AGI agent. Under CAIS, each service interacts with other services via clearly defined channels of communication, so that the system is interpretable and transparent, even though each service may be opaque. We can reason about what information is present in the inputs to infer what the service could possibly know. We could also provide access to some capability through an external resource during training, so that the service doesn't develop that capability itself.

This interpretability allows us to monitor the service -- for example, we could look at which subservices it accesses in order to make sure it isn't doing anything crazy. But what if having a human in the loop leads to unacceptable delays? Well, this would only

happen for deployed applications, where having a human in the loop seems expected, and should also be economically incentivized because it leads to better behavior. Basic AI R&D can continue to be improved autonomously without a human in the loop, so you could still see an intelligence explosion. Note that tactical tasks requiring quick reaction times probably would be delegated to AI services, but the important strategic decisions could still be left in human hands (assisted by AI services, of course).

What happens when we create AGI?

Well, it might not be valuable to create an AGI. We want to perform many different tasks, and it makes sense for these to be done by diverse services. It would not be competitive to include all capabilities in a single monolithic agent. This is analogous to how specialization of labor is a good idea for us humans.

(My opinion: It seems like the lesson of deep learning is that if you can do something end-to-end, that will work better than a structured approach. This has happened with computer vision, natural language processing, and seems to be in the process of happening with robotics. So I don't buy this -- while it seems true that we will get CAIS before AGI since structured approaches tend to be available sooner and to work with less compute, I expect that a monolithic AGI agent would outperform CAIS at most tasks once we can make one.)

That said, if we ever do build AGI, we can leverage the services from our CAIS-world in order to make it safe. We could use superintelligent security services to constrain any AGI agent that we build. For example, we could have services trained to identify long-term planning processes and to perform adversarial testing and red teaming.

Safety in the CAIS world

While CAIS suggests that we will not have AGI agents, this does not mean that we automatically get safety. We will still have AI systems that take high impact actions, and if they take even one wrong action of this sort it could be catastrophic. One way this could happen is if the system of services starts to show agentic behavior -- our standard AI safety work could apply to this scenario.

In order to ensure safety, we should have AI safety researchers figure out and codify the best development practices that need to be followed. For example, we could try to always use predictive models of human (dis)approval as a sanity check on any plan that is being enacted. We could also train AI services that can adversarially check new services to make sure they are safe.

Summary

The CAIS model suggests that before we get to a world with monolithic AGI agents, we will already have seen an intelligence explosion due to automated R&D. This reframes the problems of AI safety and has implications for what technical safety researchers should be doing.

ETA: This comment provides updates based on more discussion with Eric.

Book Review: The Structure Of Scientific Revolutions

When I hear scientists talk about Thomas Kuhn, he sounds very reasonable. Scientists have theories that guide their work. Sometimes they run into things their theories can't explain. Then some genius develops a new theory, and scientists are guided by that one. So the cycle repeats, knowledge gained with every step.

When I hear philosophers talk about Thomas Kuhn, he sounds like a madman. There is no such thing as ground-level truth! Only theory! No objective sense-data! Only theory! No basis for accepting or rejecting any theory over any other! Only theory! No scientists! Only theories, wearing lab coats and fake beards, hoping nobody will notice the charade!

I decided to read Kuhn's *The Structure Of Scientific Revolutions* in order to understand this better. Having finished, I have come to a conclusion: yup, I can see why this book causes so much confusion.

At first Kuhn's thesis appears simple, maybe even obvious. I found myself worrying at times that he was knocking down a straw man, although of course we have to read the history of philosophy backwards and remember that Kuhn may already be in the water supply, so to speak. He argues against a simplistic view of science in which it is merely the gradual accumulation of facts. So Aristotle discovered a few true facts, Galileo added a few more on, then Newton discovered a few more, and now we have very many facts indeed.

In this model, good science cannot disagree with other good science. You're either wrong – as various pseudoscientists and failed scientists have been throughout history, positing false ideas like "the brain is only there to cool the blood" or "the sun orbits the earth". Or you're right, your ideas are enshrined in the Sacristry Of Settled Science, and your facts join the accumulated store that passes through the ages.

Simple-version-of-Kuhn says this isn't true. Science isn't just facts. It's *paradigms* – whole ways of looking at the world. Without a paradigm, scientists wouldn't know what facts to gather, how to collect them, or what to do with them once they had them. With a paradigm, scientists gather and process facts in the ways the paradigm suggests ("normal science"). Eventually, this process runs into a hitch – apparent contradictions, or things that don't quite fit predictions, or just a giant ugly mess of epicycles. Some genius develops a new paradigm ("paradigm shift" or "scientific revolution"). Then the process begins again. Facts can be accumulated within a paradigm. And many of the facts accumulated in one paradigm can survive, with only slight translation effort, into a new paradigm. But scientific progress is the story of one relatively-successful and genuinely-scientific effort giving way to a different and contradictory relatively-successful and genuinely-scientific effort. It's the story of scientists constantly tossing out one another's work and beginning anew.

This gets awkward because paradigms look a lot like facts. The atomic theory – the current paradigm in a lot of chemistry – looks a lot like the fact "everything is made of atoms and molecules". But this is only the iceberg's tip. Once you have atomic theory, chemistry starts looking a lot different. Your first question when confronted with an unknown chemical is "what is the molecular structure?" and you have pretty good

ideas for how to figure this out. You are not particularly interested in the surface appearance of chemicals, since you know that iron and silver can look alike but are totally different elements; you may be much more interested in the weight ratio at which two chemicals react (which might seem to the uninitiated like a pretty random and silly thing to care about). If confronted with a gas, you might ask things like "which gas is it?" as opposed to thinking all gases are the same thing, or wondering what it would even mean for two gases to be different. You can even think things like "this is a mixture of two different types of gas" without agonizing about how a perfectly uniform substance can be a mixture of anything. If someone asks you "How noble and close to God would say this chemical sample is?" you can tell them that this is not really a legitimate chemical question, unless you mean "noble" in the sense of the noble gases. If someone tells you a certain chemical is toxic because toxicity is a fundamental property of its essence, you can tell them that no, it probably has to do with some reaction it causes or fails to cause with chemicals in the body. And if someone tells you that a certain chemical has changed into a different chemical because it got colder, you can tell them that cold might have done something to it, it might even have caused it to react with the air or something, but chemicals don't change into other chemicals in a fundamental way just because of the temperature. None of these things are obvious. All of them are hard-won discoveries.

A field without paradigms looks like the STEM supremacist's stereotype of philosophy. There are all kinds of different schools – Kantians, Aristotelians, Lockeans – who all disagree with each other. There may be progress within a school – some Aristotelian may come up with a really cool new Aristotelian way to look at bioethics, and all the other Aristotelians may agree that it's great – but the field as a whole does not progress. People will talk past one another; the Aristotelian can go on all day about the telos of the embryo, but the utilitarian is just going to ask what the hell a telos is, why anyone would think embryos have one, and how many utils the embryo is bringing people. "Debates" between the Aristotelian and the utilitarian may not be literally impossible, but they are going to have to go all the way to first principles, in a way that never works. Kuhn interestingly dismisses these areas as "the fields where people write books" – if you want to say anything, you might as well address it to a popular audience for all the good other people's pre-existing knowledge will do you, and you may have to spend hundreds of pages explaining your entire system from the ground up. He throws all the social sciences in this bin – you may read Freud, Skinner, and Beck instead of Aristotle, Locke, and Kant, but it's the same situation.

A real science is one where everyone agrees on a single paradigm. Newtonianism and Einsteinianism are the same kind of things as Aristotelianism and utilitarianism; but in 1850, everybody believed the former, and in 1950, the latter.

I got confused by this – is Aristotelian philosophy a science? Would it be one if the Aristotelians forced every non-Aristotelian philosopher out of the academy, so that 100% of philosophers fell in line behind Aristotle? I *think* Kuhn's answer to this is that it's telling that Aristotelians haven't been able to do this (at least not lately); either Aristotle's theories are too weak, or philosophy too intractable. But all physicists unite behind Einstein in a way that all philosophers cannot behind Aristotle. Because of this, all physicists mean more or less the same thing when they talk about "space" and "time", and they can work together on explaining these concepts without constantly arguing to each other about what they mean or whether they're the right way to think about things at all (and a Newtonian and Einsteinian would *not* be able to do this with each other, any more than an Aristotelian and utilitarian).

So how does science settle on a single paradigm when other fields can't? Is this the part where we admit it's because science has objective truth so you can just settle questions with experiments?

This is very much not that part. Kuhn doesn't think it's anywhere near that simple, for a few reasons.

First, there is rarely a single experiment that one paradigm fails and another passes. Rather, there are dozens of experiments. One paradigm does better on some, the other paradigm does better on others, and everyone argues over which ones should or shouldn't count.

For example, one might try to test the Copernican vs. Ptolemaic worldviews by observing the parallax of the fixed stars over the course of a year. Copernicus predicts it should be visible; Ptolemy predicts it shouldn't be. It isn't, which means either the Earth is fixed and unmoving, or the stars are unutterably unimaginably immensely impossibly far away. Nobody expected the stars to be that far away, so advantage Ptolemy. Meanwhile, the Copernicans posit far-off stars in order to save their paradigm. What looked like a test to select one paradigm or the other has turned into a wedge pushing the two paradigms even further apart.

What looks like a decisive victory to one side may look like random noise to another. Did you know weird technologically advanced artifacts are sometimes found encased in rocks that our current understanding of geology says are millions of years old? Creationists have no trouble explaining those – the rocks are much younger, and the artifacts were probably planted by nephilim. Evolutionists have no idea how to explain those, and default to things like "the artifacts are hoaxes" or "the miners were really careless and a screw slipped from their pocket into the rock vein while they were mining". I'm an evolutionist and I agree the artifacts are probably hoaxes or mistakes, even when there is no particular evidence that they are. Meanwhile, probably creationists say that some fossil or other incompatible with creationism is a hoax or a mistake. But that means the "find something predicted by one paradigm but not the other, and then the failed theory comes crashing down" oversimplification doesn't work. Find something predicted by one paradigm but not the other, and often the proponents of the disadvantaged paradigm can – and should – just shrug and say "whatever".

In 1870, flat-earther Samuel Rowbotham performed a series of experiments to show the Earth could not be a globe. In the most famous, he placed several flags miles apart along a perfectly straight canal. Then he looked through a telescope and was able to see all of them in a row, even though the furthest should have been hidden by the Earth's curvature. Having done so, he concluded the Earth was flat, and the spherical-earth paradigm debunked. Alfred Wallace (more famous for pre-empting Darwin on evolution) took up the challenge, and showed that the bending of light rays by atmospheric refraction explained Rowbotham's result. It turns out that light rays curve downward at a rate equal to the curvature of the Earth's surface! Luckily for Wallace, refraction was already a known phenomenon; if not, it would have been the same kind of wedge-between-paradigms as the Copernicans having to change the distance to the fixed stars.

It is all nice and well to say "Sure, it *looks* like your paradigm is right, but once we adjust for this new idea about the distance to the stars / the refraction of light, the evidence actually supports my paradigm". But the supporters of old paradigms can do that too! The Ptolemaics are rightly mocked for adding epicycle after epicycle until

their system gave the right result. But to a hostile observer, positing refraction effects that exactly counterbalance the curvature of the Earth sure looks like adding epicycles. At some point a new paradigm will win out, and its "epicycles" will look like perfectly reasonable adjustments for reality's surprising amount of detail. And the old paradigm will lose, and its "epicycles" will look like obvious kludges to cover up that it never really worked. Before that happens…well, good luck.

Second, two paradigms may not even address or care about the same questions.

Let's go back to utilitarianism vs. Aristotelianism. Many people associate utilitarianism with the trolley problem, which is indeed a good way to think about some of the issues involved. It might be tempting for a utilitarian to think of Aristotelian ethics as having some different answer to the trolley problem. Maybe it does, I don't know. But Aristotle doesn't talk about how he would solve whatever the 4th-century BC equivalent of the trolley problem was. He talks more about "what is the true meaning of justice?" and stuff like that. While you can twist Aristotle into having an opinion on trolleys, he's not really optimizing for that. And while you can make utilitarianism have some idea what the true meaning of justice is, it's not really optimized for that either.

An Aristotelian can say their paradigm is best, because it does a great job explicating all the little types and subtypes of justice. A utilitarian can say *their* paradigm is best, because it does a great job telling you how to act in various contrived moral dilemmas.

It's actually even worse than this. The closest thing I can think of to an ancient Greek moral dilemma is the story of Antigone. Antigone's uncle declares that her traitorous dead brother may not be buried with the proper rites. Antigone is torn between her duty to obey her uncle, and her desire to honor her dead brother. Utilitarianism is…not really designed for this sort of moral dilemma. Is ignoring her family squabbles and trying to cure typhus an option? No?

But then utilitarianism's problems are deeper than just "comes to a different conclusion than ancient Greek morals would have". The utilitarian's job isn't to change the ancient Greek's mind about the answer to a certain problem. It's to convince him to stop caring about basically all the problems he cares about, and care about different problems instead.

Third, two paradigms may disagree on what kind of answers are allowed, or what counts as solving a problem.

Kuhn talks about the 17th century "dormitive potency" discourse. Aristotle tended to explain phenomena by appealing to essences; trees grew because it was "in their nature" to grow. Descartes gets a bad rap for inventing dualism, but this is undeserved – what he was really doing was inventing the concept of "matter" as we understand it, a what-you-see-is-what-you-get kind of stuff with no hidden essences that responds mechanically to forces (and once you have this idea, you naturally need some other kind of stuff to be the mind). With Cartesian matter firmly in place, everyone made fun of Aristotle for thinking he had "solved" the "why do trees grow?" question by answering "because it is in their nature", and this climaxed with the playwright Moliere portraying a buffoonish doctor who claimed to have discovered how opium put people to sleep – it was because it had a dormitive potency!

In Aristotle's view of matter, saying "because it's their essence" successfully answers questions like "why do trees grow?". The Cartesian paradigm forbade this kind of answer, and so many previously "solved" problems like why trees grow became

mysterious again – a step backwards, sort of. For Descartes, you were only allowed to answer questions if you could explain how purely-mechanical matter smashing against other purely-mechanical matter in a billiard-ball-like way could produce an effect; a more virtuous and Descartes-aware doctor explained opium's properties by saying opium corpuscles must have a sandpaper-like shape that smooths the neurons!

Then Newton discovered gravity and caused an uproar. Gravity posits no corpuscles jostling other corpuscles. It sounds almost Aristotelian: "It is the nature of matter to attract other matter". Newton was denounced as trying to smuggle occultism into science. How much do you discount a theory for having occult elements? If some conception of quantum theory predicts the data beautifully, but says matter behaves differently depending on whether someone's watching it or not, is that okay? What if it says that a certain electron has a 50% chance of being in a certain place, full stop, and there is no conceivable explanation for which of the two possibilities is realized, and you're not even allowed to ask the question? What if my explanation for dark matter is "invisible gremlins"? How do you figure out when you need to relax your assumptions about what counts as science, versus when somebody is just cheating?

A less dramatic example: Lavoisier's theory of combustion boasts an ability to explain why some substances gain weight when burned; they are absorbing oxygen from the air. A brilliant example of an anomaly explained, which proves the superiority of combustion theory to other paradigms that cannot account for the phenomenon? No – "things shouldn't randomly gain weight" comes to us as a principle of the chemical revolution of which Lavoisier was a part:

> In the seventeenth century, [an explanation of weight gain] seemed unnecessary to most chemists. If chemical reactions could alter the volume, color, and texture of the ingredients, why should they not alter weight as well? Weight was not always taken to be the measure of quantity of matter. Besides, weight-gain on roasting remained an isolated phenomenon. Most natural bodies (eg wood) lose weight on roasting as the phlogiston theory was later to say they should.

In previous paradigms, weight gain wasn't even an anomaly to be explained. It was just a perfectly okay thing that might happen. It's only within the constellation of new methods and rules we learned around Lavoisier's time, that Lavoisier's theories solved anything at all.

So how *do* scientists ever switch paradigms?

Kuhn thinks it's kind of an ugly process. It starts with exasperation; the old paradigm is clearly inadequate. Progress is stagnating.

> Awareness [of the inadequacy of geocentric astronomy] did come. By the thirteenth century Alfonso X could proclaim that if God had consulted him when creating the universe, he would have received good advice. In the sixteenth century, Copernicus' coworker, Domenico da Novara, held that no system so cumbersome and inaccurate as the Ptolemaic had become could possibly be true of nature. And Copernicus himself wrote in the Preface to the De Revolutionibus that the astronomical tradition he inherited had finally created only a monster.

Then someone proposes a new paradigm. In its original form, it is woefully underspecified, bad at matching reality, and only beats the old paradigm in a few test cases. For whatever reason, a few people jump on board. Sometimes the new paradigm is simply more mathematically elegant, more beautiful. Other times it's petty things, like a Frenchman invented the old paradigm and a German the new one,

and you're German. Sometimes it's just that there's nothing better. These people gradually expand the new paradigm to cover more and more cases. At some point, the new paradigm explains things a little better than the old paradigm. Some of its predictions are spookily good. The old paradigm is never conclusively debunked. But the new paradigm now has enough advantages that more and more people hop on the bandwagon. Gradually the old paradigm becomes a laughingstock, people forget the context in which it ever made sense, and it is remembered only as a bunch of jokes about dormitive potency.

But now that it's been adopted and expanded and reached the zenith of its power, *this* is the point at which we can admit it's objectively better, right?

For a better treatment of this question than I can give, see Samzdat's Science Cannot Count To Red. But my impression is that Kuhn is not really willing to say this. I think he is of the "all models are wrong, some are useful" camp, thinks of paradigms as models, and would be willing to admit a new paradigm may be more useful than an old one.

Can we separate the fact around which a paradigm is based (like "the Earth orbits the sun") from the paradigm itself (being a collection of definitions of eg "planet" and "orbit", ways of thinking, mathematical methods, and rules for what kind of science will and won't be accepted)? And then say the earth factually orbits the sun, and the paradigm is just a useful tool that shouldn't be judged objectively? I think Kuhn's answer is that facts cannot be paradigm-independent. A medieval would not hear "the Earth orbits the sun" and hear the same claim we hear (albeit, in his view wrong). He would, for example, interpret it to mean the Earth was set in a slowly-turning crystal sphere with the sun at its center. Then he might ask – where does the sphere intersect the Earth? How come we can't see it? Is Marco Polo going to try to travel to China and then hit a huge invisible wall halfway across the Himalayas? And what about gravity? My understanding is the Ptolemaics didn't believe in gravity as we understand it at all. They believed objects had a natural tendency to seek the center of the universe. So if the sun is more central, why isn't everything falling into the sun? To a medieval the statement "the Earth orbits the sun" has a bunch of common-sense disproofs everywhere you look. It's only when attached to the rest of the Copernican paradigm that it starts to make sense.

This impresses me less than it impresses Kuhn. I would say "if you have many false beliefs, then true statements may be confusing in that they seem to imply false statements – but true statements are still objectively true". Perhaps I am misunderstanding Kuhn's argument here; the above is an amalgam of various things and not something Kuhn says outright in the book. But whatever his argument, Kuhn is not really willing to say that there are definite paradigm-independent objective facts, at least not without a lot of caveats.

So where *is* the point at which we admit some things are objectively true and that's what this whole enterprise rests on?

Kuhn only barely touches on this, in the last page of the book:

> Anyone who has followed the argument this far will nevertheless feel the need to ask why the evolutionary process should work. What must nature, including man, be like in order that science be possible at all? Why should scientific communities be able to reach a firm consensus unattainable in other fields? Why should consensus endure across one paradigm change after another? And why should

> paradigm change invariably produce an instrument more perfect in any sense than those known before? From one point of view those questions, excepting the first, have already been answered. But from another they are as open as they were when this essay began. It is not only the scientific community that must be special. The world of which that community is a part must also possess quite special characteristics, and we are no closer than we were at the start to knowing what these must be. That problem— What must the world be like in order that man may know it?— was not, however, created by this essay. On the contrary, it is as old as science itself, and it remains unanswered. But it need not be answered in this place.

At this point I lose patience. Kuhn is no longer being thought-provoking, he's being disingenuous. IT'S BECAUSE THERE'S AN OBJECTIVE REALITY, TOM. YOU DON'T HAVE TO BE SO COY ABOUT IT. "OHHHHH, WHAT COULD POSSIBLY EXPLAIN WHY SCIENCE BEHAVES THE WAY IT WOULD IF OBJECTIVE REALITY EXISTS, NOBODY WILL EVER KNOW, LET'S JUST NEVER ANSWER IT". Get a life.

Honestly this decreases my trust in some of what's come before. Maybe he wrote all those sections about incommensurable paradigms because paradigms really are that incommensurable. Or maybe it's because he thinks he's playing some kind of ridiculous game where the first person to admit the existence of objective reality loses.

II.

A lot of the examples above are mine, not Kuhn's. Some of them even come from philosophy or other nonscientific fields. Shouldn't I have used the book's own examples?

Yes. But one of my big complaints about this book is that, for a purported description of How Science Everywhere Is Always Practiced, it really just gives five examples. Ptolemy/Copernicus on astronomy. Alchemy/Dalton on chemistry. Phlogiston/Lavoisier on combustion. Aristotle/Galileo/Newton/Einstein on motion. And ???/Franklin/Coulomb on electricity.

It doesn't explain any of the examples. If you don't already know what Coulomb's contribution to electricity is and what previous ideas he overturned, you're out of luck. And don't try looking it up in a book either. Kuhn says that all the books have been written by people so engrossed in the current paradigm that they unconsciously jam past scientists into it, removing all evidence of paradigm shift. This made parts of the book a little beyond my level, since my knowledge of Coulomb begins and ends with "one amp per second".

Even saying Kuhn has five examples is giving him too much credit. He usually brings in one of his five per point he's trying to make, meaning that you never get a really full view of how any of the five examples exactly fit into his system.

And all five examples are from physics. Kuhn says at the beginning that he wished he had time to talk about how his system fits biology, but he doesn't. He's unsure whether any of the social sciences are sciences at all, and nothing else even gets mentioned. This means we have to figure out how Kuhn's theory fits everything from scattershot looks at the history of electricity and astronomy and a few other things. This is pretty hard. For example, consider three scientific papers I've looked at on this blog recently:

– Cipriani, Ioannidis, et al perform a meta-analysis of antidepressant effect sizes and find that although almost all of them seem to work, amitriptyline works best.

– Ceballos, Ehrlich, et al calculate whether more species have become extinct recently than would be expected based on historical background rates; after finding almost 500 extinctions since 1900, they conclude they definitely have.

– Terrell et al examine contributions to open source projects and find that men are more likely to be accepted than women when adjusted for some measure of competence they believe is appropriate, suggesting a gender bias.

What paradigm is each of these working from?

You could argue that the antidepressant study is working off of the "biological psychiatry" paradigm, a venerable collection of assumptions that can be profitably contrasted with other paradigms like psychoanalysis. But couldn't a Hippocratic four-humors physician of a thousand years ago done the same thing? A meta-analysis of the effect sizes of various kinds of leeches for depression? Sure, leeches are different from antidepressants, but it doesn't look like the belief in biological psychiatry is affecting anything about the research other than the topic. And although the topic is certainly important, Kuhn led me to expect something more profound than that. Maybe the paradigm is evidence-based-medicine itself, the practice of doing RCTs and meta-analyses on things? I think this is a stronger case, but a paradigm completely divorced from the content of what it's studying is exactly the sort of weird thing that makes me wish Kuhn had included more than five examples.

As for the extinction paper, surely it can be attributed to some chain of thought starting with Cuvier's catastrophism, passing through Lyell, and continuing on to the current day, based on the idea that the world has changed dramatically over its history and new species can arise and old ones disappear. But is that "the" paradigm of biology, or ecology, or whatever field Ceballos and Lyell are working in? Doesn't it also depend on the idea of species, a different paradigm starting with Linnaeus and developed by zoologists over the ensuing centuries? It look like it dips into a bunch of different paradigms, but is not wholly within any.

And the open source paper? Is "feminism" a paradigm? But surely this is no different than what would be done to investigate racist biases in open source. Or some right-winger looking for anti-Christian biases in open source. Is the paradigm just "looking for biases in things?"

What about my favorite trivial example, looking both ways when you cross the street so you don't get hit by a bus? Is it based on a paradigm of motorized transportation? Does it use assumptions like "buses exist" and "roads are there to be crossed"? Was there a paradigm shift between the bad old days of looking one way before crossing, and the exciting new development of looking both ways before crossing? Is this really that much more of a stretch than calling looking for biases in things a paradigm?

Outside the five examples Kuhn gives from the physical sciences, identifying paradigms seems pretty hard – or maybe too easy. Is it all fractal? Are there overarching paradigms like atomic theory, and then lower-level paradigms like organic chemistry, and then tiny subsubparadigms like "how we deal with this one organic compound"? Does every scientific experiment use lots of different paradigms from different traditions and different levels? This is the kind of thing I wish Kuhn's book answered instead of just talking about Coulumb and Copernicus over and over again.

III.

In conclusion, all of this is about predictive coding.

It's the same thing. Perception getting guided equally by top-down expectations and bottom-up evidence. Oh, I know what you're thinking. "There goes Scott again, seeing predictive coding in everything". And yes. But also, Kuhn does everything short of come out and say "When you guys get around to inventing predictive coding, make sure to notice that's what I was getting at this whole time."

Don't believe me? From the chapter *Anomaly And The Emergence Of Scientific Discovery* (my emphasis, and for "anomaly", read "surprisal"):

> The characteristics common to the three examples above are characteristic of all discoveries from which new sorts of phenomena emerge. Those characteristics include: the previous awareness of anomaly, the gradual and simultaneous emergence of both observational and conceptual recognition, and the consequent change of paradigm categories and procedures often accompanied by resistance.
>
> **There is even evidence that these same characteristics are built into the nature of the perceptual process itself.** In a psychological experiment that deserves to be far better known outside the trade, Bruner and Postman asked experimental subjects to identify on short and controlled exposure a series of playing cards. Many of the cards were normal, but some were made anomalous, e.g., a red six of spades and a black four of hearts. Each experimental run was constituted by the display of a single card to a single subject in a series of gradually increased exposures. After each exposure the subject was asked what he had seen, and the run was terminated by two successive correct identifications.
>
> Even on the shortest exposures many subjects identified most of the cards, and after a small increase all the subjects identified them all. For the normal cards these identifications were usually correct, but the anomalous cards were almost always identified, without apparent hesitation or puzzlement, as normal. The black four of hearts might, for example, be identified as the four of either spades or hearts. Without any awareness of trouble, it was immediately fitted to one of the conceptual categories prepared by prior experience. One would not even like to say that the subjects had seen something different from what they identified. With a further increase of exposure to the anomalous cards, subjects did begin to hesitate and to display awareness of anomaly. Exposed, for example, to the red six of spades, some would say: That's the six of spades, but there's something wrong with it— the black has a red border. Further increase of exposure resulted in still more hesitation and confusion until finally, and sometimes quite suddenly, most subjects would produce the correct identification without hesitation. Moreover, after doing this with two or three of the anomalous cards, they would have little further difficulty with the others. A few subjects, however, were never able to make the requisite adjustment of their categories. Even at forty times the average exposure required to recognize normal cards for what they were, more than 10 per cent of the anomalous cards were not correctly identified. And the subjects who then failed often experienced acute personal distress. One of them exclaimed: "I can't make the suit out, whatever it is. It didn't even look like a card that time. I don't know what color it is now or whether it's a spade or a heart. I'm not even sure now what a spade looks like. My God!" In the next section we shall occasionally see scientists behaving this way too.

> Either as a metaphor or **because it reflects the nature of the mind**, that psychological experiment provides a wonderfully simple and cogent schema for the process of scientific discovery.

And from *Revolutions As Changes Of World-View*:

> **Surveying the rich experimental literature from which these examples are drawn makes one suspect that something like a paradigm is prerequisite to perception itself. What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see. In the absence of such training there can only be, in William James's phrase, "a bloomin' buzzin' confusion." In recent years several of those concerned with the history of science have found the sorts of experiments described above immensely suggestive.**

If you can read those paragraphs and honestly still think I'm just just irrationally reading predictive coding into a perfectly innocent book, I have nothing to say to you.

I think this is my best answer to the whole "is Kuhn denying an objective reality" issue. If Kuhn and the predictive coding people are grasping at the same thing from different angles, then both shed some light on each other. I think I understand the way that predictive coding balances the importance of pre-existing structures and categories with a preserved belief in objectivity. If Kuhn is trying to use something like the predictive coding model of the brain processing information to understand the way the scientific community as a whole processes it, then maybe we can import the same balance and not worry about it as much.

Alignment Newsletter #39

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

Happy New Year!

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter.

Highlights

Constructing Unrestricted Adversarial Examples with Generative Models *(Yang Song et al)*: This paper predates the unrestricted adversarial examples challenge (AN #24) and shows how to generate such unrestricted adversarial examples using generative models. As a reminder, most adversarial examples research is focused on finding imperceptible perturbations to existing images that cause the model to make a mistake. In contrast, unrestricted adversarial examples allow you to find *any* image that humans will reliably classify a particular way, where the model produces some other classification.

The key idea is simple -- train a GAN to generate images in the domain of interest, and then create adversarial examples by optimizing an image to simultaneously be "realistic" (as evaluated by the generator), while still being misclassified by the model under attack. The authors also introduce another term into the loss function that minimizes deviation from a randomly chosen noise vector -- this allows them to get diverse adversarial examples, rather than always converging to the same one.

They also consider a "noise-augmented" attack, where in effect they are running the normal attack they have, and then running a standard attack like FGSM or PGD afterwards. (They do these two things simultaneously, but I believe it's nearly equivalent.)

For evaluation, they generate adversarial examples with their method and check that humans on Mechanical Turk reliably classify the examples as a particular class. Unsurprisingly, their adversarial examples "break" all existing defenses, including the certified defenses, though to be clear existing defenses assume a different threat model where an adversarial example must be an imperceptible perturbation to one of a known set of images. You could imagine doing something similar by taking the imperceptible-perturbation attacks and raise the value of ϵ until it is perceptible -- but in this case the generated images are much less realistic.

Rohin's opinion: This is the clear first thing to try with unrestricted adversarial examples, and it seems to work reasonably well. I'd love to see whether adversarial training with these sorts of adversarial examples works as a defense against both this attack and standard imperceptible-perturbation attacks. In addition, it would be interesting to see if humans could direct or control the search for unrestricted adversarial examples.

Technical AI alignment

Technical agendas and prioritization

Why I expect successful alignment *(Tobias Baumann)*: This post gives three arguments that we will likely solve the narrow alignment problem of having an AI system do what its operators intend it to do. First, advanced AI systems may be developed in such a way that the alignment problem doesn't even happen, at least as we currently conceive of it. For example, under the comprehensive AI services model, there are many different AI services that are superintelligent at particular tasks that can work together to accomplish complex goals, but there isn't a single unified agent to "align". Second, if it becomes obvious that alignment will be a serious problem, then we will devote a lot of resources to tackling the problem. We already see reward hacking in current systems, but it isn't sufficiently dangerous yet to merit the application of a lot of resources. Third, we have already come up with some decent approaches that seem like they could work.

Rohin's opinion: I generally agree with these arguments and the general viewpoint that we will probably solve alignment in this narrow sense. The most compelling argument to me is the second one, that we will eventually devote significant resources to the problem. This does depend on the crux that we see examples of these problems and how they could be dangerous before it is too late.

I also agree that it's much less clear whether we will solve other related problems, such as how to deal with malicious uses of AI, issues that arise when multiple superintelligent AI systems aligned with different humans start to compete, and how to ensure that humans have "good" values. I don't know if this implies that *on the margin* it is more useful to work on the related problems. It could be that these problems are so hard that there is not much that we can do. (I'm neglecting importance of the problem here.)

Integrative Biological Simulation, Neuropsychology, and AI Safety *(Gopal Sarma et al)*: This paper argues that we can make progress on AI capabilities and AI safety through integrative biological simulation, that is, a composite simulation of all of the processes involved in neurons that allow us to simulate brains. In the near future, such simulations would be limited to simple organisms like Drosophila, but even these organisms exhibit behavior that we find hard to replicate today using our AI techniques, especially at the sample efficiency that the organisms show. On the safety side, even such small brains share many architectural features with human brains, and so we might hope that we could discover neuroscience-based methods for value learning that generalize well to humans. Another possibility would be to create test suites (as in AI Safety Gridworlds) for simulated organisms.

Rohin's opinion: I don't know how hard it would be to create integrative biological simulations, but it does strike me as very useful if we did have them. If we had a complete mechanistic understanding of how intelligence happens in biological brains (in the sense that we can simulate them), the obvious next step would be to understand *how* the mechanistic procedures lead to intelligence (in the same way that we currently try to understand why neural nets work). If we succeed at this, I would expect to get several insights into intelligence that would translate into significant progress in AI. However, I know very little about biological neurons and brains so take this with many grains of salt.

On the value learning side, it would be a good test of inverse reinforcement learning to see how well it could work on simple organisms, though it's not obvious what the ground truth is. I do want to note that this is specific to inverse reinforcement learning

-- other techniques depend on uniquely human characteristics, like the ability to answer questions posed by the AI system.

Agent foundations

Robust program equilibrium *(Caspar Oesterheld)*: In a prisoner's dilemma where you have access to an opponent's source code, you can hope to achieve cooperation by looking at how the opponent would perform against you. Naively, you could simply simulate what the opponent would do given your source code, and use that to make your decision. However, if your opponent also tries to simulate you, this leads to an infinite loop. The key idea of this paper is to break the infinite loop by introducing a small probability of guaranteed cooperation (without simulating the opponent), so that eventually after many rounds of simulation the recursion "bottoms out" with guaranteed cooperation. They explore what happens when applying this idea to the equivalents of FairBot/Tit-for-Tat strategies when you are simulating the opponent.

Preventing bad behavior

Penalizing Impact via Attainable Utility Preservation *(Alex Turner)*: This post and the linked paper present [Attainable Utility Preservation] (AN #25) more simply. There are new experiments that show that AUP works on some of the AI Safety Gridworlds even when using a set of *random* utility functions, and compares this against other methods of avoiding side effects.

Rohin's opinion: While this is easier to read and understand, I think there are important points in the original post that do not come across, so I would recommend reading both. In particular, one of my core takeaways from AUP was that convergent instrumental subgoals could be avoided by penalizing *increases* in attainable utilities, and I don't think that comes across as well in this paper. This is the main thing that makes AUP different, and it's what allows it to avoid disabling the off switch in the Survival gridworld.

The fact that AUP works with random rewards is interesting, but I'm not sure it will generalize to realistic environments. In these gridworlds, there is usually a single thing that the agent is not supposed to do. It's very likely that several of the random rewards will care about that particular thing, which means that the AUP penalty will apply, so as long as full AUP would have solved the problem, AUP with random rewards would probably also solve it. However, in more realistic environments, there are many different things that the agent is supposed to avoid, and it's not clear how big a random sample of reward functions needs to be in order to capture all of them. (However, it does seem reasonably likely that if the reward functions are "natural", you only need a few of them to avoid convergent instrumental subgoals.)

Adversarial examples

Constructing Unrestricted Adversarial Examples with Generative Models *(Yang Song et al)*: Summarized in the highlights!

Near-term concerns

Fairness and bias

Learning Not to Learn: Training Deep Neural Networks with Biased Data *(Byungju Kim et al)*

AI strategy and policy

AI Index 2018 Report *(Yoav Shoham et al)*: Lots of data about AI. The report highlights how AI is global, the particular improvement in natural language understanding over the last year, and the limited gender diversity in the classroom. We also see the expected trend of huge growth in AI, both in terms of interest in the field as well as in performance metrics.

AI Now 2018 Report *(Meredith Whittaker et al)*: See Import AI

Sequence introduction: non-agent and multiagent models of mind

A typical paradigm by which people tend to think of themselves and others is as *consequentialist agents*: entities who can be usefully modeled as having beliefs and goals, who are then acting according to their beliefs to achieve their goals.

This is often a useful model, but it doesn't quite capture reality. It's a bit of a fake framework. Or in computer science terms, you might call it a leaky abstraction.

An abstraction in the computer science sense is a simplification which tries to hide the underlying details of a thing, letting you think in terms of the simplification rather than the details. To the extent that the abstraction actually succeeds in hiding the details, this makes things a lot simpler. But sometimes the abstraction inevitably leaks, as the simplification fails to predict some of the actual behavior that emerges from the details; in that situation you need to actually know the underlying details, and be able to think in terms of them.

Agent-ness being a leaky abstraction is not exactly a novel concept for Less Wrong; it has been touched upon several times, such as in Scott Alexander's Blue-Minimizing Robot Sequence. At the same time, I do not think that it has been quite fully internalized yet, and that many foundational posts on LW go wrong due to being premised on the assumption of humans being agents. In fact, I would go as far as to claim that this is the biggest flaw of the original Sequences: they were attempting to explain many failures of rationality as being due to cognitive biases, when in retrospect it looks like understanding cognitive biases doesn't actually make you substantially more effective. But if you are implicitly modeling humans as goal-directed agents, then cognitive biases is the most natural place for irrationality to emerge from, so it makes sense to focus the most on there.

Just knowing that an abstraction leaks isn't enough to improve your thinking, however. To do better, you need to know about the actual underlying details to get a better model. In this sequence, I will aim to elaborate on various tools for thinking about minds which look at humans in more granular detail than the classical agent model does. Hopefully, this will help us better get past the old paradigm.

My model of what I think our subagents looks like draws upon a number of different sources, including neuroscience, psychotherapy and meditation, so in the process of sketching out my model I will be covering a number of them in turn. To give you a rough idea of what I'm trying to do, here's a summary of some upcoming content.

Published posts:

(*Note: this list may not always be fully up to date; see the sequence index for actively maintained version)*

Book summary: Consciousness and the Brain. One of the fundamental building blocks of much of consciousness research, is that of Global Workspace Theory (GWT). This could be described as a component of a multiagent model, focusing on the way in which different agents exchange information between one another. One elaboration of

GWT, which focuses on how it might be implemented in the brain, is the Global Neuronal Workspace (GNW) model in neuroscience. Consciousness in the Brain is a 2014 book that summarizes some of the research and basic ideas behind GNW, so summarizing the main content of that book looks like a good place to start our discussion and for getting a neuroscientific grounding before we get more speculative.

Building up to an IFS model. One theoretical approach for modeling humans as being composed of interacting parts is that of Internal Family Systems. In my experience and that of several other people in the rationalist community, it's very effective for this purpose. However, having its origins in therapy, its theoretical model may seem rather unscientific and woo-y. This personally put me off the theory for a long time, as I thought that it sounded fake, and gave me a strong sense of "my mind isn't split into parts like that".

In this post, I construct a mechanistic sketch of how a mind might work, drawing on the kinds of mechanisms that have already been demonstrated in contemporary machine learning, and then end up with a model that pretty closely resembles the IFS one.

Subagents, introspective awareness, and blending. In this post, I extend the model of mind that I've been building up in previous posts to explain some things about change blindness, not knowing whether you are conscious, forgetting most of your thoughts, and mistaking your thoughts and emotions as objective facts, while also connecting it with the theory in the meditation book The Mind Illuminated.

Subagents, akrasia, and coherence in humans. We can roughly describe coherence as the property that, if you become aware that there exists a more optimal strategy for achieving your goals than the one that you are currently executing, then you will switch to that better strategy. For a subagent theory of mind, we would like to have some explanation of when exactly the subagents manage to be collectively coherent (that is, change their behavior to some better one), and what are the situations in which they fail to do so.

My conclusion is that we are capable of changing our behaviors on occasions when the mind-system as a whole puts sufficiently high probability on the new behavior being better, when the new behavior is not being blocked by a particular highly weighted subagent (such as an IFS-style protector) that puts high probability on it being bad, and when we have enough slack in our lives for any new behaviors to be evaluated in the first place. Akrasia is subagent disagreement about what to do.

Integrating disagreeing subagents. In the previous post, I suggested that akrasia involves subagent disagreement - or in other words, different parts of the brain having differing ideas on what the best course of action is. The existence of such conflicts raises the question, how does one resolve them?

In this post I discuss various techniques which could be interpreted as ways of resolving subagents disagreements, as well as some of the reasons for why this doesn't always happen.

Subagents, neural Turing machines, thought selection, and blindspots. In my summary of *Consciousness and the Brain*, I briefly mentioned that one of the functions of consciousness is to carry out artificial serial operations; or in other words, implement a production system (equivalent to a Turing machine) in the brain.

While I did not go into very much detail about this model in the post, I've used it in

later articles. For instance, in *Building up to an Internal Family Systems model*, I used a toy model where different subagents cast votes to modify the contents of consciousness. One may conceptualize this as equivalent to the production system model, where different subagents implement different production rules which compete to modify the contents of consciousness.

In this post, I flesh out the model a bit more, as well as applying it to a few other examples, such as emotion suppression, internal conflict, and blind spots.

Subagents, trauma, and rationality. This post interprets the appearance of subagents as emerging from *unintegrated memory networks*, and argues that the presence of these is a matter of degree. There's a continuous progression of fragmented (dissociated) memory networks giving arise to increasingly worse symptoms as the degree of fragmentation grows. The continuum goes from everyday procrastination and akrasia on the "normal" end, to disrupted and dysfunctional beliefs on the middle, and conditions like clinical PTSD, borderline personality disorder, and dissociative identity disorder on the severely traumatized end.

I also argue that emotional work and exploring one's past traumas in order to heal them, is necessary for effective instrumental and epistemic rationality.

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

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

Book summary: Unlocking the Emotional Brain. Written by the psychotherapists Bruce Ecker, Robin Ticic and Laurel Hulley, *Unlocking the Emotional Brain* claims to offer a neuroscience-grounded, comprehensive model of how effective therapy works. In so doing, it also happens to formulate its theory in terms of belief updating, helping explain how the brain models the world and what kinds of techniques allow us to actually change our minds. Its discussion and models are closely connected to the models about internal conflict and belief revision that are discussed in previous posts, particularly "integrating disagreeing subagents".

A mechanistic model of meditation. Meditation has been claimed to have all kinds of transformative effects on the psyche, such as improving concentration ability, healing trauma, cleaning up delusions, allowing one to track their subconscious strategies, and making one's nervous system more efficient. However, an explanation for why and how exactly this would happen has typically been lacking. This makes people reasonably skeptical of such claims.

In this post, I want to offer an explanation for one kind of a mechanism: meditation increasing the degree of a person's introspective awareness, and thus leading to increasing psychological unity as internal conflicts are detected and resolved.

A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble. Insight meditation, enlightenment, what's that all about?

The sequence of posts starting from this one is my personal attempt at answering that question. It seeks to:

- Explain what kinds of implicit assumptions build up our default understanding of reality and how those assumptions are subtly flawed.
- Point out aspects from our experience whose repeated observation will update those assumptions, and explain how this may cause psychological change in someone who meditates.
- Explain how the so-called "three characteristics of existence" of Buddhism - impermanence, no-self and unsatisfactoriness - are all interrelated and connected with each other in a way that is connected to the previously discussed topics in the sequence.

Farther out (sketched out but not as extensively planned/written yet)

The game theory of rationality and cooperation in a multiagent world. Multi-agent models have a natural connection to Elephant in the Brain -style dynamics: our brains doing things for purposes of which we are unaware. Furthermore, there can be strong incentives to continue systematic self-deception and *not* integrate conflicting beliefs. For instance, if a mind has subagents which think that specific beliefs are dangerous to hold or express, then they will work to suppress subagents holding that belief from coming into conscious awareness.

"Dangerous beliefs" might be ones that touch upon political topics, but they might also be ones of a more personal nature. For instance, someone may have an identity as being "good at X", and then want to rationalize away any contradictory evidence - including evidence suggesting that they were wrong on a topic related to X. Or it might be something even more subtle.

These are a few examples of how rationality work has to happen on two levels at once: to debug some beliefs (individual level), people need to be in a community where holding various kinds of beliefs is *actually safe* (social level). But in order for the community to be safe for holding those beliefs (social level), people within the community also need to work on themselves so as to deal with their own subagents that would cause them to attack people with the "wrong" beliefs (individual level). This kind of work also seems to be necessary for fixing "politics being the mind-killer" and collaborating on issues such as existential risk across sharp value differences; but the need to carry out the work on many levels at once makes it challenging, especially since the current environment incentivizes many (sub)agents to sabotage any attempt at this.

(This topic area is also related to that stuff Valentine has been saying about Omega.)

This sequence is part of research done for, and supported by, the Foundational Research Institute.

Visualizing the power of multiple step selection processes in JS: Galton's bean machine

This is a linkpost for http://www.gwern.net/docs/statistics/order/beanmachine-multistage/index.html

Book Recommendations: An Everyone Culture and Moral Mazes

Epistemic Status: Casual

I highly recommend An Everyone Culture, by Robert Kegan, and Moral Mazes, by Robert Jackall, as companion books on business culture. *Moral Mazes* is an anthropological study of the culture and implicit ethics of a few large corporations, and is an eye-opening illustration of the problems that arise in those corporations. *An Everyone Culture* is an introduction to the idea of a "deliberately developmental organization", an attempt to fix those problems, plus some case studies of companies that implemented "deliberately developmental" practices.

The basic problem that both books observe in corporate life is that *everybody in a modern office is trying to conceal their failures and present a misleadingly positive impression of themselves to their employers and coworkers.*

This leads to lost productivity.

For instance:

- The longer one tries to cover up a mistake, the costlier it will be to fix it.
- The less accurately credit is allocated for success or failure, the harder it will be to incentivize good work.
- The more employees misinform their bosses, the worse-informed the bosses' decisions will be.
- The more people are concerned with maintaining appearances, the less cognitive capacity they will have for productivity and creativity.
- The more unacceptable it is to acknowledge "personal" concerns (emotions, physical health, intrinsic motivation or lack thereof), the harder it is to fix productivity problems that arise from "personal" problems.

Moral Mazes basically takes the view that the Protestant work ethic really died in the mid-to-late nineteenth century, when an American economy defined by small business owners and freelance professionals was replaced by an economy defined by larger firms and the rise of the managerial profession. The Protestant work ethic declared that hard work, discipline, and honesty would bring success. The "managerial work ethic" holds that a good employee has quite different "virtues" — things like

- ability to play politics
- loyalty & willing to subordinate oneself to one's manager
- "flexibility" (the opposite of stubbornness — not holding strong individual opinions)

To give an outside example, the author of "The Western Elite from a Chinese Perspective" was coming from a "Protestant work ethic" culture of hard work (though not, of course, actually Protestant) and encountering the "managerial work ethic" culture of American office politics.

Moral Mazes relies on the author's observations and interviews with managers. I'm sure it's not a fully objective portrayal — perhaps the author selected the most

damning quotes, and perhaps the most disgruntled and cynical managers were the most willing to talk. But the picture the book gives is of a culture where:

- rank is everything — contradicting your boss, especially in public, is career suicide, and deference to superiors is expected
- beyond a certain minimum floor of competence, objective job performance doesn't determine career success, political skill does
- "credit flows upwards, details flow downwards" — higher-rank managers take credit for work done by their subordinates, and the higher-rank you are, the fewer object-level details you concern yourself with
- mistakes and bad decisions are *reliably* concealed; then, when the inevitable catastrophe happens, whoever's politically vulnerable takes the fall
- managers are tested for their "flexibility" — someone with strong opinions about the best engineering decisions or with rigid ethical principles will not rise far in their career

If you watch *The Marvelous Mrs. Maisel*, Joel Maisel's job at the plastics company is a classic example of the managerial work ethic; he's basically a professional sycophant. He's burned out and unmotivated, and he leaves to "find himself" as a comedian, but quickly realizes he has no talent at comedy either. Instead, working in his father's garment business, he comes to life again. He learns the nitty-gritty of the factory floor, the accounting, the machines, the seamstresses and their personal needs and strengths and weaknesses. It's a beautiful illustration of the difference between fake work and real work.

An Everyone Culture's prescription for the problems of deception, sycophancy, and stagnation in conventional companies is complex, but I'd summarize it as follows: creating a culture where *everyone talks about mistakes and improvements*, and *where the personal/professional boundaries are broken down.*

This sounds vaguely cultish and shocking, and indeed, the companies profiled (like Bridgewater) are often described as cults. Kegan acknowledges that their practices are outside most of our comfort zones, but believes that nothing *inside* the range of what we think of as a normal workplace will solve workplace dysfunctions.

What distinguishes the companies profiled in the book is a *lot* of talk, about issues that would ordinarily be considered too "personal" for work. When someone makes a mistake, a DDO looks for the root cause, as you would in a kaizen system, but it won't stop there — people will also ask what *personal or psychological* issue caused the mistake. Does this person have a tendency towards overconfidence that they need to work on? Were they afraid of looking bad? Do they need to learn to consider others' feelings more?

It's vulnerable to be laid bare in this way, but, at least in the ideal of a DDO, *everyone* does it, from the interns to the CEO, to the point that people internalize that *having flaws and a personal life is nothing to hide*. Some people would find this horrifically intrusive, but others find it a relief.

I've never worked in a DDO, but I think I might like it; with enough mandated transparency, I'd be forced to override the temptation to hide flaws and make myself look better, and could focus better on actually *doing* good work.

The cost, of course, is *way more communication* about seemingly non-work-related things. You'd be processing personal stuff with coworkers all the time. The hope is that this is actually cheaper than the costs of the bad decisions made when you don't have

enough honest communication, but it’s an empirical matter whether that works out in practice, and the authors don’t have data so far.

CDT=EDT=UDT

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

Epistemic status: I no longer endorse the particular direction this post advocates, though I'd be excited if someone figured out something that seems to work. I still endorse most of the specific observations.

So... what's the deal with counterfactuals?

Over the past couple of years, I've been writing about the CDT=EDT perspective. I've now organized those posts into a sequence for easy reading.

I call CDT=EDT a "perspective" because it is a way of consistently answering questions about what counterfactuals are and how they work. At times, I've argued strongly that it is the *correct* way. That's basically because:

- it has been the *only* coherent framework I put any stock in (more for lack of other proposals for dealing with logical counterfactuals than for an abundance of bad ones);
- there *are* strong arguments for it, *if* you're willing to make certain assumptions;
- it would be awfully nice to settle this whole question of counterfactual reasoning and move on. CDT=EDT is in a sense the most boring possible answer, IE that all approaches we've thought of are essentially equivalent and there's no hope for anything better.

However, recently I've realized that there's a perspective which unifies *even more* approaches, while being *less boring* (more optimistic about counterfactual reasoning helping us to do well in decision-theoretic problems). It's been right in front of me the whole time, but I was blind to it due to the way I factored the problem of formulating decision theory. It suggests a research direction for making progress in our understanding of counterfactuals; I'll try to indicate some open curiosities of mine by the end.

Three > Two

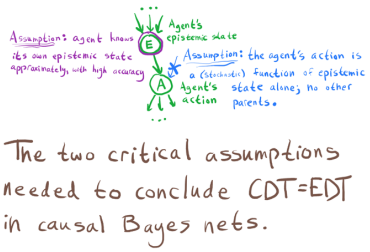
The claim I'll be elaborating on in this post is, essentially, that the framework in Jessica Taylor's post about memoryless cartesian environments is better than the CDT=EDT way of thinking. You'll have to read the post to get the full picture if you haven't, but to briefly summarize: if we formalize decision problems in a framework which Jessica Taylor calls "memoryless cartesian environments" (which we can call "memoryless POMDPs" if we want to be closer to academic CS/ML terminology), reasoning about anthropic uncertainty in a certain way (via the self-indication assumption, SIA for short) makes it possible for CDT to behave like UDT.

The result there is sometimes abbreviated as UDT=CDT+SIA, although UDT⊂CDT+SIA is more accurate, because the optimal UDT policies are a subset of the policies which CDT+SIA can follow. This is because UDT has self-coordination power which CDT+SIA lacks. (We could say UDT=CDT+SIA+coordination, but unfortunately "coordination" lacks a snappy three-letter acronym. Or, to be even more pedantic, we could say that

UDT1.0 = CDT+SIA, and UDT1.1 = CDT+SIA+coordination. (The difference between 1.0 and 1.1 is, after all, the presence of global policy coordination.)) [EDIT: This isn't correct. See Wei Dai's comment.]

Caspar Oesterheld commented on that post with an analogous EDT+SSA result. SSA (the self-sampling assumption) is one of the main contenders beside SIA for correct anthropic reasoning. Caspar's comment shows that we can think of the correct anthropics as a function of your preference between CDT and EDT. So, we could say that CDT+SIA = EDT+SSA = UDT1.0; or, CDT=EDT=UDT for short. [EDIT: As per Wei Dai's comment, the equation "CDT+SIA = EDT+SSA = UDT1.0" is really not correct due to differing coordination strengths; as he put it, UDT1.0 > EDT+SSA > CDT+SIA.]

My CDT=EDT view came from being pedantic about how decision problems are represented, and noticing that when you're pedantic, it becomes awfully hard to drive a wedge between CDT and EDT; you've got to do things which are strange enough that it becomes questionable whether it's a fair comparison between CDT and EDT. However, I didn't notice the extent to which my "being very careful about the representation" was really *insisting that bayes nets are the proper representation*.



*(Aside: Bayes nets which are representing decision problems are usually called **influence diagrams** rather than Bayes nets. I think this convention is silly; why do we need a special term for that?)*

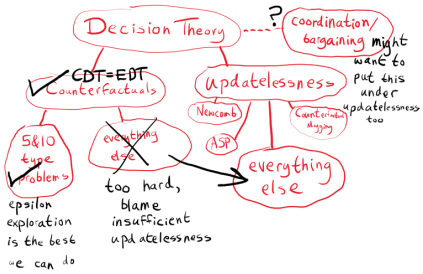
It is rather curious that LIDT also illustrated CDT=EDT-style behavior. It is part of what made me feel like CDT=EDT was a convergent result of many different approaches, rather than noticing its reliance on certain Bayes-net formulations of decision problems. Now, I instead find it to be curious and remarkable that logical induction seems to think as if the world were made of bayes nets.

If CDT=EDT comes from insisting that decision problems are represented as Bayes nets, CDT=EDT=UDT is the view which comes from insisting that decision problems be represented as memoryless cartesian environments. At the moment, this just seems like a better way to be pedantic about representation. It unifies three decision theories instead of two.

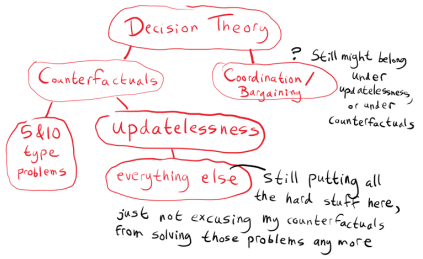
Updatelessness Doesn't Factor Out

In fact, I thought about Jessica's framework frequently, but I didn't think of it as an objection to my CDT=EDT way of thinking. I was blind to this objection because I thought (logical-)counterfactual reasoning and (logically-)updateless reasoning could be dealt with as separate problems. The claim was not that CDT=EDT-style decision-making did well, but rather, that any decision problem where it performed poorly could be analyzed as a case where updateless reasoning is needed in order to do well.

I let my counterfactual reasoning be simple, blaming all the hard problems on the difficulty of logical updatelessness.



Once I thought to question this view, it seemed very likely wrong. The Dutch Book argument for CDT=EDT seems closer to the true justification for CDT=EDT reasoning than the Bayes-net argument, but the Dutch Book argument is a dynamic consistency argument. I know that CDT and EDT both violate dynamic consistency, in general. So, why pick on one special type of dynamic consistency violation which CDT can illustrate but EDT cannot? In other words, the grounds on which I can argue CDT=EDT seem to point more directly to UDT instead.



What about all those arguments for CDT=EDT?

Non-Zero Probability Assumptions

I've noted before that each argument I make for CDT=EDT seems to rely on an assumption that actions have non-zero probability. I leaned heavily on an assumption of epsilon exploration, although one could also argue that all actions must have non-zero probability on different grounds (such as the implausibility of knowing so much about what you are going to do that you can completely rule out any action, before you've made the decision). Focusing on cases where we have to assign probability zero to some action was a big part of finally breaking myself of the CDT=EDT view and moving to the CDT=EDT=UDT view.

(I was almost broken of the view about a year ago by thinking about the XOR blackmail problem, which has features in common with the case I'll consider now; but, it didn't stick, perhaps because the example doesn't actually force actions to have probability zero and so doesn't point so directly to where the arguments break down.)

Consider the transparent Newcomb problem with a perfect predictor:

Transparent Newcomb. *Omega runs a perfect simulation of you, in which you face two boxes, a large box and a small box. Both boxes are made of transparent glass. The small box contains $100, while the large one contains $1,000. In the Simulation, Omega gives you the option of either taking both boxes or only taking the large box. If Omega predicts that you will take only one box, then Omega puts you in this situation for real. Otherwise, Omega gives the real you the same decision, but with the large box empty. You find yourself in front of two full boxes. Do you take one, or two?*

Apparently, since Omega is a perfect predictor, we are forced to assign probability zero to one-boxing even if we follow a policy of epsilon-exploring. In fact, if you implement epsilon-exploration by refusing to take any action which you're very confident you'll take (you have a hard-coded response: if **P("I do action X")>1-epsilon**, do anything but **X**), which is how I often like to think about it, then ***you are forced to 2-box in transparent Newcomb***. I was *expecting* CDT=EDT type reasoning to 2-box (at which point I'd say "but we can fix that by being updateless"), but this is a *really weird reason* to 2-box.

Still, that's not in itself an argument against CDT=EDT. Maybe the rule that we can't take actions we're overconfident in is at fault. The argument against CDT=EDT style counterfactuals in this problem is that the agent should expect that if it 2-boxes, then it won't ever be in the situation to begin with; at least, not in the *real* world. As discussed somewhat in the happy dance problem, this breaks important properties that you might want out of conditioning on conditionals. (There are some interesting consequences of this, but they'll have to wait for a different post.) More importantly for the CDT=EDT question, this can't follow from evidential conditioning, or learning about consequences of actions through epsilon-exploration, or any other principles in the CDT=EDT cluster. So, there would at least have to be other principles in play.

A very natural way of dealing with the problem is to represent the agent's uncertainty about whether it is in a simulation. If you think you might be in Omega's simulation, observing a full box doesn't imply certainty about your own action anymore, or even about whether the box is really full. This is exactly how you deal with the problem in memoryless cartesian environments. But, if we are willing to do this here, we might as well think about things in the memoryless cartesian framework all over the place. This contradicts the CDT=EDT way of thinking about things in lots of problems where updateless reasoning gives different answers than updatefull reasoning, such as counterfactual mugging, rather than only in cases where some action has probability zero.

(I should actually say "problems where updateless reasoning gives different answers than *non-anthropic* updateful reasoning", since the whole point here is that updateful reasoning *can* be consistent with updateless reasoning so long as we take anthropics into account in the right way.)

I also note that trying to represent this problem in bayes nets, while possible, is very awkward and dissatisfying compared to the representation in memoryless cartesian environments. You could say I shouldn't have gotten myself into a position where this felt like significant evidence, but, reliant on Bayes-net thinking as I was, it did.

Ok, so, looking at examples which force actions to have probability zero made me revise my view even for cases where actions all have non-zero probability. So again, it makes sense to ask: but what about the arguments in favor of CDT=EDT?

Bayes Net Structure Assumptions

The argument in the bayes net setting makes some assumptions about the structure of the Bayes net, illustrated earlier. Where do those go wrong?

In the Bayes net setting, observations are represented as parents of the epistemic state (which is a parent of the action). To represent the decision conditional on an observation, we condition on the observation being true. This stops us from putting some probability on our observations being false due to us being in a simulation, as we do in the memoryless cartesian setup.

In other words: the CDT=EDT setup makes it impossible to update on something and still have rational doubt in it, which is what we need to do in order to have an updateful DT act like UDT.

There's likely *some* way to fix this while keeping the Bayes-net formalism. However, memoryless cartesian environments model it naturally.

Question: how can we model memoryless cartesian environments in Bayes nets? Can we do this in a way such that the CDT=EDT theorem applies (making the CDT=EDT way of thinking compatible with the CDT=EDT=UDT way of thinking)?

CDT Dutch Book

What about the Dutch-book argument for CDT=EDT? I'm not quite sure how this one plays out. I need to think more about the setting in which the Dutch-book can be carried out, especially as it relates to anthropic problems and anthropic Dutch-books.

Learning Theory

I said that I think the Dutch-book argument gets closer to the real reason CDT=EDT seems compelling than the Bayes-net picture does. Well, although the Dutch Book argument against CDT gives a crisp justification of a CDT=EDT view, I felt the learning-theoretic intuitions which lead me to formulate the dutch book are closer to the real story. It doesn't make sense to ask an agent to have good counterfactuals in any single situation, because the agent may be ignorant about how to reason about the situation. However, any errors in counterfactual reasoning which result in observed consequences predictably differing from counterfactual expectations should eventually be corrected.

I'm still in the dark about how this argument connects to the CDT=EDT=UDT picture, just as with the Dutch-book argument. I'll discuss this more in the next section.

Static vs Dynamic

A big update in my thinking recently has been to cluster frameworks into "static" and "dynamic", and ask how to translate back and forth between static and dynamic versions of particular ideas. Classical decision theory has a strong tendency to think in terms of statically given decision problems. You could say that the epistemic problem of figuring out what situation you're in is assumed to factor out: decision theory deals

only with what to do once you're in a particular situation. On the other hand, learning theory deals with more "dynamic" notions of rationality: rationality-as-improvement-over-time, rather than an absolute notion of perfect performance. (For our purposes, "time" includes logical time; even in a single-shot game, you can learn from relevantly similar games which play out in thought-experiment form.)

This is a messy distinction. Here are a few choice examples:

Static version: Dutch-book and money-pump arguments.

Dynamic version: Regret bounds.

Dutch-book arguments rely on the idea that you shouldn't *ever* be able to extract money from a rational gambler without a chance of losing it instead. Regret bounds in learning theory offer a more relaxed principle, that you can't ever extract *too much* money (for some notion of "too much" given by the particular regret bound). The more relaxed condition is more broadly applicable; Dutch-book arguments only give us the probabilistic analog of logical consistency properties, whereas regret bounds give us inductive learning.

Static: Probability theory.

Dynamic: Logical induction.

In particular, the logical induction criterion gives a notion of regret which implies a large number of nice properties. Typically, the difference between logical induction and classical probability theory is framed as one of logical omniscience vs logical uncertainty. The static-vs-dynamic frame instead sees the critical difference as one of rationality in a static situation (where it makes sense to think about perfect reasoning) vs learning-theoretic rationality (where it doesn't make sense to ask for perfection, and instead, one thinks in terms of regret bounds).

Static: Bayes-net decision theory (either CDT or EDT as set up in the CDT=EDT argument).

Dynamic: LIDT.

As I mentioned before, the way LIDT seems to naturally reason as if the world were made of Bayes nets now seems like a curious coincidence rather than a convergent consequence of correct counterfactual conditioning. I would like a better explanation of why this happens. Here is my thinking so far:

- Logical induction lacks a way to question its perception. As with the Bayes-net setup used in the CDT=EDT argument, to observe something is to think that thing is true. There is not a natural way for logical induction to reason anthropically, especially for information which comes in through the traders thinking longer. If one of the traders calculates digits of π and bets accordingly, this information is simply known by the logical inductor; how can it entertain the possibility that it's in a simulation and the trader's calculation is being modified by Omega?
- Logical induction knows its own epistemic state to within high accuracy, as is assumed in the Bayes-net CDT=EDT theorem.
- LIDT makes the action a function of the epistemic state alone, as required.

There's a lot of formal work one could do to try to make the connection more rigorous (and look for places where the connection breaks down!).

Static: UDT.

Dynamic: ???

The problem of logical updatelessness has been a thorn in my side for some time now. UDT is a good reply to a lot of decision-theoretic problems when they're framed in a probability-theoretic setting, but moving to a logically uncertain setting, it's unclear how to apply UDT. UDT requires a fixed prior, whereas logical induction gives us a picture in which logical uncertainty is fundamentally about how to revise beliefs as you think longer.

The main reason the static-vs-dynamic idea has been a big update for me is that I realized that a lot of my thinking has been aimed at turning logical uncertainty into a "static" object, to be able to apply UDT. I haven't even posted about most of those ideas, because they haven't lead anywhere interesting. Tsvi's post on thin logical priors is definitely an example, though. I now think this type of approach is likely doomed to failure, because the dynamic perspective is simply superior to the static one.

The interesting question is: how do we translate UDT to a dynamic perspective? How do we learn updateless behavior?

For all its flaws, taking the dynamic perspective on decision theory feels like something asymptotic decision theory got right. I have more to say about what ADT does right and wrong, but perhaps it is too much of an aside for this post.

A general strategy we might take to approach that question is: how do we translate individual things which UDT does right into learning-theoretic desiderata? (This may be more tractable than trying to translate the UDT optimality notion into a learning-theoretic desideratum whole-hog.)

Static: Memoryless Cartesian decision theories (CDT+SIA or EDT+SSA).

Dynamic: ???

The CDT=EDT=UDT perspective on counterfactuals is that we can approach the question of learning logically updateless behavior by thinking about the learning-theoretic version of anthropic reasoning. How do we learn which observations to take seriously? How do we learn about what to expect supposing we *are* being fooled by a simulation? Some optimistic speculation on that is the subject of the next section.

We Have the Data

Part of why I was previously very pessimistic about doing any better than the CDT=EDT-style counterfactuals was that we *don't have any data* about counterfactuals, almost by definition. How are we supposed to learn what to counterfactually expect? We only observe the real world.

Consider LIDT playing transparent Newcomb with a perfect predictor. Its belief that it will 1-box in cases where it sees that the large box is full must converge to 100%,

because it only ever sees a full box in cases where it does indeed 1-box. Furthermore, the expected utility of 2-boxing can be anything, since it will never see cases where it sees a full box and 2-boxes. This means I can make LIDT 1-box by designing my LI to think 2-boxing upon seeing a full box will be catastrophically bad: I simply include a trader with high initial wealth who bets it will be bad. Similarly, I can make LIDT 2-box whenever it sees the full box by including a trader who bets 2-boxing will be great. Then, the LIDT will never see a full box except on rounds where it is going to epsilon-explore into 1-boxing.

(The above analysis depends on details of how epsilon exploration is implemented. If it is implemented via the probabilistic chicken-rule, mentioned earlier, making the agent explore whenever it is very confident about which action it takes, then the situation gets pretty weird. Assume that LIDT is epsilon-exploring pseudorandomly instead.)

LIDT's confidence that it 1-boxes whenever it sees a full box is jarring, because I've just shown that I can make it either 1-box or 2-box depending on the underlying LI. Intuitively, an LIDT agent who 2-boxes upon seeing the full box should not be near-100% confident that it 1-boxes.

The problem is that the cases where LIDT sees a full box and 2-boxes are all counterfactual, since Omega is a perfect predictor and doesn't show us a full box unless we in fact 1-box. LIDT doesn't learn from counterfactual cases; the version of the agent in Omega's head is shut down when Omega is done with it, and never reports its observations back to the main unit.

(The LI *does* correctly learn the *mathematical fact* that its algorithm 2-boxes when input observations of a full box, but, this does not help it to have the intuitively correct expectations when Omega feeds it false sense-data.)

In the terminology of The Happy Dance Problem, LIDT isn't learning the right observation-counterfactuals: the predictions about what action it takes given different possible observations. However, ***we have the data:*** the agent *could* simulate itself under alternative epistemic conditions, and train its observation-counterfactuals on what action it in fact takes in those conditions.

Similarly, the action-counterfactuals are wrong: LIDT can believe anything about what happens when it 2-boxes upon seeing a full box. Again, ***we have the data:*** LI can observe that on rounds when it is mathematically true that the LIDT agent would have 2-boxed upon seeing a full box, it doesn't get the chance. This knowledge simply isn't being "plugged in" to the decision procedure in the right way. Generally speaking, an agent can observe the real consequences of counterfactual actions, because (1) the counterfactual action is a mathematical fact of what the agent does under a counterfactual observation, and (2) the important effects of this counterfactual action occur in the real world, which we can observe directly.

This observation makes me much more optimistic about learning interesting counterfactuals. Previously, it seemed like *by definition* there would be no data from which to learn the correct counterfactuals, other than the (EDTish) requirement that they should match the actual world for actions actually taken. Now, it seems like I have not one, but *two* sources of data: the observation-counterfactuals can be simulated outright, and the action-counterfactuals can be trained on what actually happens when counterfactual actions are taken.

I haven't been able to plug these pieces together to get a working counterfactual-learning algorithm yet. It might be that I'm still missing a component. But ... it *really* feels like there should be something here.

Supervising strong learners by amplifying weak experts

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://arxiv.org/pdf/1810.08575.pdf

Abstract

Many real world learning tasks involve complex or hard-to-specify objectives, and using an easier-to-specify proxy can lead to poor performance or misaligned behavior. One solution is to have humans provide a training signal by demonstrating or judging performance, but this approach fails if the task is too complicated for a human to directly evaluate. We propose Iterated Amplification, an alternative training strategy which progressively builds up a training signal for difficult problems by combining solutions to easier subproblems. Iterated Amplification is closely related to Expert Iteration (Anthony et al., 2017; Silver et al., 2017b), except that it uses no external reward function. We present results in algorithmic environments, showing that Iterated Amplification can efficiently learn complex behaviors.

Tomorrow's AI Alignment Forum sequences post will be 'AI safety without goal-directed behavior' by Rohin Shah, in the sequence on Value Learning.

The next post in this sequence on Iterated Amplification will be 'AlphaGo Zero and capability amplification', by Paul Christiano, on Tuesday 8th January.

Does anti-malaria charity destroy the local anti-malaria industry?

The usual argument against foreign aid to Africa is that randomly giving tons of free goods (such as food) *ruins local producers*; and when at some later moment the charity goes out of fashion (or decides to target a different part of Africa), the local situation becomes *even worse than before*, because the local producers have gone out of business. In addition, it hurts the local people psychologically to know that *any* local business, no matter how successful it could otherwise have been, can at any moment be destroyed by a well-meaning foreign charity.

Recently I heard the same argument made about anti-malaria nets recommended by GiveWell. If I understand it correctly, the donated nets put local net producers out of business (increasing local poverty and dependence on foreign aid), and the estimated number of lives saved is misleading (because in the alternative scenario, the same people could have been saved by locally produced nets).

I have one specific question, and one more general concern.

The specific question... well, I know *nothing* about the anti-malaria industry in Africa. It exists, I assume. But *quantitatively* -- how many nets it produces, how many nets it stops producing because it is pushed out of the market by GiveWell, whether the nets are of comparable quality, what is the best estimate of the scenario with no foreign aid compared to the scenario with foreign aid -- I have no idea. I supposed some of this was *already discussed* by some effective altruists, so I would love to hear the summary.

The meta concern is the following: I find the argument of foreign goods disrupting local market plausible. But seems to me that the problem is with *high variance* (one year a ton of goods, the very next year nothing), not with foreign goods *per se*. Because, anytime a country participates in foreign trade, the local producers of the stuff that is being imported, are pushed out of business. But we have the law of comparative advantages saying that in global, this is a *good* thing, for both countries. (Or to put it differently, trade *sanctions* are typically used as a punishment, not as a reward.) I worry that at some moment, the "stop destroying African economy by your disruptive aid" argument becomes effectively "stop trading with Africa", and I am not sure where exactly to draw that line.

Alignment Newsletter #40

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter.

The Alignment Forum sequences have started again! As a reminder, treat them as though I had highlighted them.

Highlights

Reframing Superintelligence: Comprehensive AI Services as General Intelligence *(Eric Drexler)*: This is a huge document; rather than summarize it all in this newsletter, I wrote up my summary in this post. For this newsletter, I've copied over the description of the model, but left out all of the implications and critiques.

The core idea is to look at the pathway by which we will develop general intelligence, rather than assuming that at some point we will get a superintelligent AGI agent. To predict how AI will progress in the future, we can look at how AI progresses currently -- through research and development (R&D) processes. AI researchers consider a problem, define a search space, formulate an objective, and use an optimization technique in order to obtain an AI system, called a service, that performs the task.

A service is an AI system that delivers bounded results for some task using bounded resources in bounded time. Superintelligent language translation would count as a service, even though it requires a very detailed understanding of the world, including engineering, history, science, etc. Episodic RL agents also count as services.

While each of the AI R&D subtasks is currently performed by a human, as AI progresses we should expect that we will automate these tasks as well. At that point, we will have automated R&D, leading to recursive technological improvement. This is not recursive self-improvement, because the improvement comes from R&D services creating improvements in basic AI building blocks, and those improvements feed back into the R&D services. All of this should happen before we get any powerful AGI agents that can do arbitrary general reasoning.

Rohin's opinion: I'm glad this has finally been published -- it's been informing my views for a long time now. I broadly buy the general view put forward here, with a few nitpicks that you can see in the post. I really do recommend you read at least the post -- that's just the *summary* of the report, so it's full of insights, and it should be interesting to technical safety and strategy researchers alike.

I'm still not sure how this should affect what research we do -- techniques like preference learning and recursive reward modeling seem applicable to CAIS as well, since they allow us to more accurately specify what we want each individual service to do.

Technical AI alignment

Iterated amplification sequence

Supervising strong learners by amplifying weak experts *(Paul Christiano)*: This was previously covered in AN #30, I've copied the summary and opinion. This paper introduces iterated amplification, focusing on how it can be used to define a training signal for tasks that humans cannot perform or evaluate, such as designing a transit system. The key insight is that humans are capable of decomposing even very difficult tasks into slightly simpler tasks. So, in theory, we could provide ground truth labels for an arbitrarily difficult task by a huge tree of humans, each decomposing their own subquestion and handing off new subquestions to other humans, until questions are easy enough that a human can directly answer them.

We can turn this into an efficient algorithm by having the human decompose the question only once, and using the current AI system to answer the generated subquestions. If the AI isn't able to answer the subquestions, then the human will get nonsense answers. However, as long as there are questions that the human + AI system can answer but the AI alone cannot answer, the AI can learn from the answers to those questions. To reduce the reliance on human data, another model is trained to predict the decomposition that the human performs. In addition, some tasks could refer to a large context (eg. evaluating safety for a specific rocket design), so they model the human as being able to access small pieces of the context at a time.

They evaluate on simple algorithmic tasks like distance between nodes in a graph, where they can program an automated human decomposition for faster experiments, and there is a ground truth solution. They compare against supervised learning, which trains a model on the ground truth answers to questions (which iterated amplification does not have access to), and find that they can match the performance of supervised learning with only slightly more training steps.

Rohin's opinion: This is my new favorite post/paper for explaining how iterated amplification works, since it very succinctly and clearly makes the case for iterated amplification as a strategy for generating a good training signal. I'd recommend reading the paper in full, as it makes other important points that I haven't included in the summary.

Note that it does not explain a lot of Paul's thinking. It explains one particular training method that allows you to train an AI system with a more intelligent and informed overseer.

Value learning sequence

Will humans build goal-directed agents? *(Rohin Shah)*: The previous post argued that coherence arguments do *not* mean that a superintelligent AI must have goal-directed behavior. In this post, I consider other arguments suggesting that we'll build goal-directed AI systems.

- Since humans are goal-directed, they will build goal-directed AI to help them achieve their goals. *Reaction:* Somewhat agree, but this only shows that the human + AI system should be goal-directed, not the AI itself.

- Goal-directed AI can exceed human performance. *Reaction:* Mostly agree, but there could be alternatives that still exceed human performance.

- Current RL agents are goal-directed. *Reaction:* While the math says this, in practice this doesn't seem true, since RL agents learn from experience rather than planning over the long term.

- Existing intelligent agents are goal-directed. *Reaction:* Seems like a good reason to not build AI using evolution.

- Goal-directed agents are more interpretable and so more desirable. *Reaction:* Disagree, it seems like we're arguing that we should build goal-directed AI so that we can more easily predict that it will cause catastrophe.

AI safety without goal-directed behavior *(Rohin Shah)*: The main thrust of the second chapter of the sequence is that it is not *required* for a superintelligent AI system to be goal-directed. While there are certainly economic arguments suggesting that we will build goal-directed AI, these do not have the force of a theorem. Given the strong arguments we've developed that goal-directed AI would likely be dangerous, it seems worth exploring other options. Some possibilities are AI systems that infer and follow norms, corrigible AI, and bounded and episodic AI services.

These other possibilities can be cast in a utility-maximization framework. However, if you do that then you are once again tempted to say that you are screwed if you get the utility function slightly wrong. Instead, I would want to build these systems in such a way that the desirable properties are inherent to the way that they reason, so that it isn't even a coherent question to ask "what if we get it slightly wrong".

Problems

Imitation learning considered unsafe? *(capybaralet)*: We might hope that using imitation learning to mimic a corrigible human would be safe. However, this would involve mimicking the human's planning process. It seems fairly likely that slight errors in the imitation of this process could lead to the creation of a goal-directed planning process that does dangerous long-term optimization.

Rohin's opinion: This seems pretty similar to the problem of inner optimizers, in which while searching for a good policy for some task T on training distribution D, you end up finding a consequentialist agent that is optimizing some utility function that leads to good performance on D. That agent will have all the standard dangers of goal-directed optimization out of distribution.

Two More Decision Theory Problems for Humans *(Wei Dai)*: The first problem is that any particular human's values only make sense for the current environment. When considering different circumstances (eg. an astronomically large number of very slightly negative experiences like getting a dust speck in your eye), many people will not know how to evaluate the value of such a situation.

The second problem is that for most formalizations of values or utility functions, the values are defined relative to some way of making decisions in the world, or some ontology through which we understand the world. If this decision theory or ontology changes, it's not clear how to "transfer" the values to the new version.

Predictors as Agents

Technical agendas and prioritization

Reframing Superintelligence: Comprehensive AI Services as General Intelligence *(Eric Drexler)*: Summarized in the highlights!

Agent foundations

Failures of UDT-AIXI, Part 1: Improper Randomizing *(Diffractor)*

Preventing bad behavior

Optimization Regularization through Time Penalty *(Linda Linsefors)*

Anthropic probabilities: answering different questions

What is the answer to the question of anthropic probabilities? I've claimed before that there was no answer - anthropic decision theory (ADT) was the only way to go.

But actually, there are answers - the problem is simply that there are multiple versions of "**the** question of anthropic probabilities", each with their own answer. And what decision theory did was unambiguously select the right question (and the right answer) for the job.

Frequentist anthropic probabilities

It is much easier for humans to think in terms of frequencies, and anthropic probabilities are no exceptions. So imagine that either a small universe (with one copy of you) or a large universe (with ω copies of you) is created, with equal probability.

Then your copies will independently observe a large sequence of Ω random bits, with $2^{\Omega} >> \omega$. After that, the universe ends, and the whole experiment begins again, with a small or large universe being created again. This experiment will then be repeated a *very large* number of times, so we can coherent talk about frequencies.

Then there are three questions you might ask yourself during these experiments:

1. What proportion of my versions will be in a large universe?
2. What proportion of universes, where versions of me exist, will be large?
3. What proportion of universes, where **exact copies** of me exist, will be large?

In the limit as these experiments are run a large number of times, the answers to these questions will converge on $\omega/(1 + \omega)$, 1/2, and "it depends on how many random bits you have seen when you ask the question". In other words, the probabilities given by SIA, SSA, and FNC.

Notice there is a fourth question that we could ask to complete the three:

4. What proportion my **exact copies** will be in a large universe?

But this question will also converge to $\omega/(1 + \omega)$, ie SIA, showing how SIA is independent of the reference class as long as the reference class contains at least the exact copies.

The issues with the questions

All three questions are well-posed questions with exact and correct answers. From outside, however, there are issues with all three.

Question 2 has the perennial "reference class problem": what are you counting as "versions of me"? If we change the reference class, we change the question, and therefore its not surprising it gives a different answer.

Question 3 has the same time inconsistency that FNC has: the answer will be (predictably) different at different times, in a way that breaks probabilities = expectation of future probabilities. Again, each question is individually sound, but "exact copies of me" means different things at different times.

Question 1 has a similar time inconsistency issue when the number of identical copies changes predictably but differentially across time.

Aside from that, questions 1 and 3 are often the wrong questions to ask in decision theory. Non-identical versions can timelessly cooperate with you; identical copies may be totally opposed to you.

The advantages of decision theory

Why does decision theory perform well in anthropic contexts, giving single decisions even when there are multiple anthropic probability questions? Simply because it unambiguously selects the question that it is relevant to answer. Average utilitarians maximise their score by figuring out the universe; total utilitarians by figuring out where most of the copies are. The whole process of ADT/UDT decision-making computes a specific reference class: the reference class of all correlated decisions with your own. By automatically including precommitments, ADT/UDT resolves the fact that the class of "exact copies of me" keeps on changing. And by being explicitly a decision theory, it resolves the issue of cooperation and non-cooperation of identical and non-identical agents.

So, back when I thought "anthropic probabilities" were a single question with a single answer, the fact that ADT/UDT gave a single answer (albeit a decision one rather than a probability one) convinced me that anthropic decisions were true while anthropic probabilities were not.

But now that I've realised that there are multiple anthropic probability questions (and also that all the "paradoxes" of anthropic probabilities have non-paradoxical decision theory analogues), I'm fully content to say:

- "Yes Virginia, anthropic probabilities exist, and different anthropic probabilities are answering different anthropic questions."

Incidentally, there are far more than three questions - each of these questions can be different, depending on what time it is asked. So I'd also conclude:

- The reason that anthropic probability is so debated, is because none of the anthropic questions correspond to a simple, stable question that corresponds to an intuitive understanding of what anthropic probability actually is.

Learning with catastrophes

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

A *catastrophe* is an event so bad that we are not willing to let it happen even a single time. For example, we would be unhappy if our self-driving car *ever* accelerates to 65 mph in a residential area and hits a pedestrian.

Catastrophes present a theoretical challenge for traditional machine learning—typically there is no way to reliably avoid catastrophic behavior without strong statistical assumptions.

In this post, I'll lay out a very general model for catastrophes in which they are avoidable under much weaker statistical assumptions. I think this framework applies to the most important kinds of catastrophe, and will be especially relevant to AI alignment.

Designing practical algorithms that work in this model is an open problem. In a subsequent post I describe what I currently see as the most promising angles of attack.

Modeling catastrophes

We consider an agent A interacting with the environment over a sequence of episodes. Each episode produces a transcript τ, consisting of the agent's observations and actions, along with a reward $r \in [0, 1]$. Our primary goal is to quickly learn an agent which receives high reward. (Supervised learning is the special case where each transcripts consist of a single input and a label for that input.)

While training, we assume that we have an oracle which can determine whether a transcript τ is "catastrophic." For example, we might show a transcript to a QA analyst and ask them if it looks catastrophic. This oracle can be applied to arbitrary sequences of observations and actions, including those that don't arise from an actual episode. So training can begin before the very first interaction with nature, using only calls to the oracle.

Intuitively, a transcript should only be marked catastrophic if it satisfies two conditions:

1. The agent made a catastrophically bad decision.
2. The agent's observations are plausible: we have a right to expect the agent to be able to handle those observations.

While actually interacting with the environment, the agent cannot query the oracle—there is no time to wait for a QA engineer to review a proposed action to check if it would be catastrophic.

Moreover, if interaction with nature ever produces a catastrophic transcript, we immediately fail. The performance of an algorithm is characterized by two parameters: the probability of catastrophic failure, and the total reward assuming no catastrophic failure.

We assume that there are some policies such that no matter what nature does, the resulting transcript is *never* catastrophic.

Traditionally in RL the goal is to get as much reward as the best policy from some class C. We' slightly weaken that goal, and instead aim to do as well as the best policy from C that never makes a catastrophic decision.

Batch learning

I've described an online version of learning with catastrophes. We can also consider the batch version, where the learner is first given a large number of "training" episodes.

In the batch version, there is no penalty for catastrophes at training time, and we don't care about training error. The two performance parameters are test-time performance and test-time catastrophe probability.

The oracle

This definition depends on an oracle who determines which transcripts are catastrophic.

For weak AI systems, the oracle may be a human. But a powerful AI system might take actions which are catastrophic but which look inoffensive to a human judge, so this approach doesn't cut it.

In general, the judge should be a human+AI team which is *more* competent than the system being trained, armed with an adequate solution to the informed oversight problem.

Approach

Learning with catastrophes is straightforward *given an unlimited number of queries to the catastrophe oracle*. Given any online learning algorithm A, we can "harden" it by running the following process before the beginning of each episode (a similar process can be applied to a batch learner):

1. Search over all possible environments, running A on each one to obtain a transcript.
2. If we find a catastrophic transcript τ, then add τ to A's training data with a reward of -1, and go back to step 1.

It's easy to prove that this process converges, if A is competent: the number of times we invoke step 2 is at most the time required to learn an optimal catastrophe-free policy (plus the number of episodes).

The big problem with this algorithm is the exponential search in step 1.

Adversarial training

The most obvious approach to *tractable* learning with catastrophes is to replace step 1 with a tractable algorithm that approximates the search. This process is an example of adversarial training.

We can formalize this as a simple game between the agent A and an adversary B. The two agents alternate taking moves, with A playing the role of the agent and B playing the role of the environment. B wins if the resulting transcript is catastrophic, A wins if it is not catastrophic.

Before each episode, we play this game N times, where $N > 0$ is a hyperparameter. If B is able to win *any* of the games, then we add the resulting transcript to A's training data and start again. Once B is not able to win any games, then we move on to a "real" episode. We hope that B is good enough at the game that if it can't produce a catastrophic transcript, then the real environment can't produce a catastrophic transcript either.

More precisely, before each episode we perform the following process:

1. Set $i = 0$.
2. A and B alternate taking moves, producing a transcript τ.
3. If τ is catastrophic, we add τ to A's training data with a reward of -1, and add τ to B's training data with a reward of $+1$. Then we go back to step 1.
4. If τ is not catastrophic, we add τ to B's training data with a reward of -1.
5. If $i < N$, we increment i and go back to step 2.

I discuss this idea in more detail in my post on red teams. There are serious problems with this approach and I don't think it can work on its own, but fortunately it seems combinable with other techniques.

Conclusion

Learning with catastrophes is a very general model of catastrophic failures which avoids being obviously impossible. I think that designing competent algorithms for learning with catastrophes may be an important ingredient in a successful approach to AI alignment.

This was originally posted here on 28th May, 2016.

Tomorrow's AI Alignment sequences post will be in the sequence on Value Learning by Rohin Shah.

The next post in this sequence will be 'Thoughts on Reward Engineering' by Paul Christiano, on Thursday.

What are good ML/AI related prediction / calibration questions for 2019?

I'm trying to come up with a set of questions for self-calibration, related to AI and ML.

I've written down what I've come up with so far below. But I am principally interested in what other people come up with -- thus the question metatype -- both for questions, and for predictions for the questions.

So far I have an insufficient number of questions to produce anything like a nice calibration curve. I've also struggled with coming up with meaningful questions.

I've rot13'ed my predictions to avoid anchoring anyone. I'm pretty uncertain about most of these as point estimates however.

On Explicitly Stated ML / Systems Goals

It is (relatively) easy to determine if these are fulfilled or not. The trade-off is that they likely have little relation to AGI.

1. OpenAI succeeds in defeating top pro teams on unrestricted Dota2

OpenAI has explicitly said that they wish to beat top human teams in the MOBA Dota2. Their latest attempt to do so used self-play and familiar policy-gradient strategies on an incredibly massive scale to train, but still lost to top teams who won (relatively?) easily.

I'm also interested in people's probabilities on whether OpenAI succeeds, conditional on OpenAI not including genuine algorithmic novelty in their learning methods, although that's a harder question to define because of cloudiness around "algorithmic novelty."

My prediction: Friragl-svir creprag.

2. Tesla succeeds in a self-driving car driving coast-to-coast without intervention.

Tesla sells cars with (ostensibly) all the hardware necessary for full self-driving, and an in-house self-driving research program that uses a mix of ML and hard-coded rules. They have a goal of giving a demonstration autonomous coast-to-coast drive, although this goal has been repeatedly delayed. There is widespread skepticism both of the sensor suite in Tesla cars and of the maturity of their software.

My prediction: Gra creprag.

3. DeepMind reveals a skilled RL-trained agent for StarCraft II.

After AlphaGo, DeepMind announced that they would try to create an expert-level agent for SCII. They've released preliminary research related to this topic, although what they've revealed is far from such an agent.

My prediction: Nobhg svir creprag.

On Goals Not So Clearly Marked As Targets

1. Someone gets a score on the Winograd Schema tests above 80%.

The Winograd Schemas are a series of tests designed to test the common-sense reasoning properties of a system. Modern ML struggles to get better than random chance -- a 50% is approximately equal to random guessing, and modern state of the art gets less than 70%. (This is the best score I could find; there are several papers which claim better scores, but these deal with subsets of the Winograd Schemas as far as I can tell. [I.e., the classic "Hey, we got a better score... on a more constrained dataset."] I might be wrong about this, if I'm wrong please illuminate me.)

My prediction: Nobhg svir creprag.

2. Reinforcement Learning Starts Working

This is a bad, unclear goal. I'm not sure how to make it clearer and could use help.

There are a lot of articles about how reinforcement learning doesn't work. It generalizes incredibly poorly, and only succeeds on complex tasks like Dota2 by playing literal centuries worth of games. If some algorithm were discovered, such that one could get RL to work with kind of the same regularity that supervised learning works, that would be amazing. I'm still struggling with a way to rigorously note this. A 10x improvement in sample efficiency on the Atari suite would (probably?) fulfill this, but I'm not sure what else would. And it's quite a PITA to keep track of what the current state-of-the-art on Atari is anyhow.

My prediction: Need to get this more defined.

"The Unbiased Map"

“Long have we suffered under the tyranny of maps.

Biased maps which show topography, but not population.

Wretched maps which speak of religion, but not languages.

Divisive maps which paint with the color of Party, but not the color of economic conditions.

Dirty maps which show crop yield across the heartland, but neglect Fiber Optic Internet coverage.

Our age calls for better maps, maps free from the bias of these old maps, perfect maps.

Imagine the day of the unbiased map. The map which shows both how to get to the airport via public transit and GDP by county. The holy map demonstrating last year’s rainfall and the distribution of seminaries and rabbinical schools. The ancestral map depicting migration of immigrants and American tribes in 1491.

Don’t give me an atlas which pretends at perfection but hisses red herrings from hydra-heads. I want the real thing, a map which doesn’t end at some arbitrary border whether it be the county line, or the sphere of earth. A map which can show the world as known by the Qing Dynasty, Strabo, Majorcan Jews, and the Aztecs. A map of Elon Musk’s neurons and a map of the solar system.

Today’s maps enlighten as the Brothers Grimm, through a bundle of fairy tales. There are no ethical maps under capitalism, all of them drip with the status quo. None show me the world that should be, none provide directions to Valhalla, all show but the thin surface of Reality. And for Mankind, the surface does not satisfy!”

Buy shares in a megaproject

I think we should create a separate legal construct for megaprojects.

I recently wrote about megaproject management here. In the outline for Oxford's Major Programme Management program, the first module is described in this way:

> Develop your understanding of major programmes as a governance structure and distinctive organisational form. In the context of programme performance, consider and reflect on organisational theory and design.

The basic concept is that instead of a normal corporation, it would operate like a corporation with an expiration date and the project goals baked into the governance. We want "on-time and on-budget" to take the place of "growth and revenue" in the new organization's decision making.

- We can buy stock in corporations, which now live ~16 years and declining.
- Megaprojects like the Big Dig (16 years of construction) and the F-35 development (26 years to production) are easily in the range of normal corporations
- Currently the only way to bet for or against the megaproject is to bet on the corporations/creditors/etc.
- This would allow us to specifically invest in, or bet against, the project.
- We invest in time-limited financial assets like bonds, puts and options already.

Consider the Boring Company. If one of those projects is a terrible idea, we are stuck with adjusting our position on the company as a whole, which means the terrible project puts the others in jeopardy.

By contrast, if the megaprojects were all separate entities and the Boring Company was hired to *manage* all of them, we could then bet on Boring Company and all of the projects independently of each other.

It seems to me that this would do a pretty good job resolving a lot of the problems that plague megaprojects currently.

Future directions for narrow value learning

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

Narrow value learning is a huge field that people are already working on (though not by that name) and I can't possibly do it justice. This post is primarily a list of things that I think are important and interesting, rather than an exhaustive list of directions to pursue. (In contrast, the corresponding post for ambitious value learning did aim to be exhaustive, and I don't think I missed much work there.)

You might think that since so many people are already working on narrow value learning, we should focus on more neglected areas of AI safety. However, I still think it's worth working on because long-term safety suggests a particular subset of problems to focus on; that subset seems quite neglected.

For example, a lot of work is about how to improve current algorithms in a particular domain, and the solutions encode domain knowledge to succeed. This seems not very relevant for long-term concerns. Some work assumes that a handcoded featurization is given (so that the true reward is linear in the features); this is not an assumption we could make for more powerful AI systems.

I will speculate a bit on the neglectedness and feasibility of each of these areas, since for many of them there isn't a person or research group who would champion them whom I could defer to about the arguments for success.

The big picture

This category of research is about how you could take narrow value learning algorithms and use them to create an aligned AI system. Typically, I expect this to work by having the narrow value learning enable some form of corrigibility.

As far as I can tell, nobody outside of the AI safety community works on this problem. While it is far too early to stake a confident position one way or the other, I am slightly less optimistic about this avenue of approach than one in which we create a system that is directly trained to be corrigible.

Avoiding problems with goal-directedness. How do we put together narrow value learning techniques in a way that doesn't lead to the AI behaving like a goal-directed agent at each point? This is the problem with keeping a reward estimate that is updated over time. While reward uncertainty can help avoid some of the problems, it does not seem sufficient by itself. Are there other ideas that can help?

Dealing with the difficulty of "human values". Cooperative IRL makes the unrealistic assumption that the human knows her reward function exactly. How can we make narrow value learning systems that deal with this issue? In particular, what prevents them from updating on our behavior that's not in line with our "true values", while still letting them update on other behavior? Perhaps we could make an AI system that is always uncertain about what the true reward is, but how does this

mesh with epistemics, which suggest that you can get to arbitrarily high confidence given sufficient evidence?

Human-AI interaction

This section of research aims to figure out how to create human-AI systems that successfully accomplish tasks. For sufficiently complex tasks and sufficiently powerful AI, this overlaps with the big picture concerns above, but there are also areas to work on with subhuman AI with an eye towards more powerful systems.

Assumptions about the human. In any feedback system, the update that the AI makes on the human feedback depends on the assumption that the AI makes about the human. In Inverse Reward Design (IRD), the AI system assumes that the reward function provided by a human designer leads to near-optimal behavior in the training environment, but may be arbitrarily bad in other environments. In IRL, the typical assumption is that the demonstrations are created by a human behaving Boltzmann rationally, but recent research aims to also correct for any suboptimalities they might have, and so no longer assumes away the problem of systematic biases. (See also the discussion in Future directions for ambitious value learning.) In Cooperative IRL, the AI system assumes that the human models the AI system as approximately rational. COACH notes that when you ask a human to provide a reward signal, they provide a critique of current behavior rather than a reward signal that can be maximized.

Can we weaken the assumptions that we have to make, or get rid of them altogether? Barring that, can we make our assumptions more realistic?

Managing interaction. How should the AI system manage its interaction with the human to learn best? This is the domain of active learning, which is far too large a field for me to summarize here. I'll throw in a link to Active Inverse Reward Design, because I already talked about IRD and I helped write the active variant.

Human policy. The utility of a feedback system is going to depend strongly on the quality of the feedback given by the human. How do we train humans so that their feedback is most useful for the AI system? So far, most work is about how to adapt AI systems to understand humans better, but it seems likely there are also gains to be had by having humans adapt to AI systems.

Finding and using preference information

New sources of data. So far preferences are typically learned through demonstrations, comparisons or rankings; but there are likely other useful ways to elicit preferences. Inverse Reward Design gets preferences from a stated proxy reward function. An obvious one is to learn preferences from what people say, but natural language is notoriously hard to work with so not much work has been done on it so far, though there is some. (I'm pretty sure there's a lot more in the NLP community that I'm not yet aware of.) We recently showed that there is even preference information in the state of the world that can be extracted.

Handling multiple sources of data. We could infer preferences from behavior, from speech, from given reward functions, from the state of the world, etc. but it seems quite likely that the inferred preferences would conflict with each other. What do you do in these cases? Is there a way to infer preferences simultaneously from all the

sources of data such that the problem does not arise? (And if so, what is the algorithm implicitly doing in cases where different data sources pull in different directions?)

Acknowledging Human Preference Types to Support Value Learning talks about this problem and suggests some aggregation rules but doesn't test them. Reward Learning from Narrated Demonstrations learns from both speech and demonstrations, but they are used as complements to each other, not as different sources for the same information that could conflict.

I'm particularly excited about this line of research -- it seems like it hasn't been explored yet and there are things that can be done, especially if you allow yourself to simply detect conflicts, present the conflict to the user, and then trust their answer. (Though this wouldn't scale to superintelligent AI.)

Generalization. Current deep IRL algorithms (or deep anything algorithms) do not generalize well. How can we infer reward functions that transfer well to different environments? Adversarial IRL is an example of work pushing in this direction, but my understanding is that it had limited success. I'm less optimistic about this avenue of research because it seems like in general function approximators do not extrapolate well. On the other hand, I and everyone else have the strong intuition that a reward function should take fewer bits to specify than the full policy, and so should be easier to infer. (Though not based on Kolmogorov complexity.)

How could shares in a megaproject return value to shareholders?

In my previous post on buying shares in a megaproject, several people asked the most reasonable of questions: how would the shares return value?

A simple description of shareholder value:

Assets - Liabilities = Equity

Basic project outcome:

Budget - Costs = Surplus

So the concept is fundamentally to treat the budget as the assets column, put the costs in the liability column, and treat every dollar you are under the budget as an increase in shareholder equity. At the conclusion of the project, the shares will be bought back for that surplus in cash.

This means the price the share sells for is a direct bet on how well the project will do against the budget.

This is analogous to how investors buy in to start-ups; the investment becomes the assets and in exchange they get a share of the equity. But in the case of megaprojects, the initial investor is getting the project outcome as compensation; the equity should instead go to management as their compensation. Now this gives management the incentive to be as efficient as possible, while *also* giving them the flexibility to raise more capital in exchange for equity.

For comparison with some of the ways this is currently done, consider the fixed-price contract and the cost-plus contract. The former gives the contractor an incentive to be efficient by forcing them to hold all the risk; the latter shifts the risk back to the contractee. Fixed-price is common for things that are predictable, like services; cost-plus is common for things that are risky or require R&D to complete, like defense procurement. Megaprojects are usually of the latter sort.

Tangential but worth also addressing: why a financial asset instead of a prediction market?

- Financial assets provide direct incentives, whereas prediction markets provide information.
- Financial assets are a built-in reference class; a megaproject asset would naturally be compared against all other megaproject assets.
- Financial assets have a huge market, and there are established methods for making bets for/against/on/with them. There are no large prediction markets, so an entire market would have to be built. The problem of incentivizing the new market seems more difficult than the one of incentivizing a new type of asset.

Two More Decision Theory Problems for Humans

(This post has been sitting in my drafts folder for 6 years. Not sure why I didn't make it public, but here it is now after some editing.)

There are two problems closely related to the Ontological Crisis in Humans. I'll call them the "Partial Utility Function Problem" and the "Decision Theory Upgrade Problem".

Partial Utility Function Problem

As I mentioned in a previous post, the only apparent utility function we have seems to be defined over an ontology very different from the fundamental ontology of the universe. But even on it's native domain, the utility function seems only partially defined. In other words, it will throw an error (i.e., say "I don't know") on some possible states of the heuristical model. For example, this happens for me when the number of people gets sufficiently large, like 3^^^3 in Eliezer's Torture vs Dust Specks scenario. When we try to compute the expected utility of some action, how should we deal with these "I don't know" values that come up?

(Note that I'm presenting a simplified version of the real problem we face, where in addition to "I don't know", our utility function could also return essentially random extrapolated values outside of the region where it gives sensible outputs.)

Decision Theory Upgrade Problem

In the Decision Theory Upgrade Problem, an agent decides that their current decision theory is inadequate in some way, and needs to be upgraded. (Note that the Ontological Crisis could be considered an instance of this more general problem.) The question is whether and how to transfer their values over to the new decision theory.

For example a human might be be running a mix of several decision theories: reinforcement learning, heuristical model-based consequentialism, identity-based decision making (where you adopt one or more social roles, like "environmentalist" or "academic" as part of your identity and then make decisions based on pattern matching what that role would do in any given situation), as well as virtual ethics and deontology. If you are tempted to drop one or more of these in favor of a more "advanced" or "rational" decision theory, such as UDT, you have to figure out how to transfer the values embodied in the old decision theory, which may not even be represented as any kind of utility function, over to the new.

Another instance of this problem can be seen in someone just wanting to be a bit more consequentialist. Maybe UDT is too strange and impractical, but our native model-based consequentialism at least seems closer to being rational than the other decision procedures we have. In this case we tend to assume that the consequentialist module already has our real values and we don't need to "port" values from the other decision procedures that we're deprecating. But I'm not entirely sure this is safe, since the step going from (for example) identity-based decision making to heuristical model-based consequentialism doesn't seem *that* different from the step between heuristical model-based consequentialism and something like UDT.

No surjection onto function space for manifold X

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

[*Note: highly technical. Skip if topology is not your thing*]

In his post on formal open problems in decision theory, Scott asked whether there could exist a topological space X and a continuous surjection s from X to $C(X, I)$. Here, I is the closed unit interval $[0, 1]$ and $C(X, I)$ is the set of continuous functions from X to I.

I thought I had an argument for how $\mathbb{R} - \mathbb{N}$ could be such an X. But that argument is wrong, as I'll demonstrate in this post. Instead I will show that:

- Let X be a manifold (with or without boundary), or a union of finitely or countably many manifolds. Then there is no continuous surjective map from X to $C(X, I)$.

By "union", imagine the manifolds lying inside Euclidean space (or, more generically, inside a metric space), not necessarily disjoint, and taking their unions there. Note that there are many examples of such Xs - for example, the rationals within the reals (being countable unions of points, which are trivial manifolds).

In fact, I will show the more general:

- If X is Fréchet–Urysohn, and σ-compact, then there is no continuous surjective map from X to $C(X, I)$.

To see that this more general result implies the one above, note that manifolds are σ-compact, and that if X_i is σ-compact, it can be covered by countably many compact sets, so $\bigcup_{i \in \mathbb{N}} X_i$ can be covered by countably many sets of countably many compact sets, which is just countably many. Finally, all metric spaces are Fréchet–Urysohn.

Fréchet–Urysohn basically means "convergence of subsequences makes sense in the topology", and is not a strong restriction; indeed all first-countable spaces are Fréchet–Urysohn.

Proof

A note on topologies

Now $C(X, I)$ is well-defined as a set, but it needs a topology to discuss issues of continuity. There are three natural topologies on it: the topology of uniform convergence, the compact-open topology (which, on this set, is equal to the topology of compact convergence), and the topology of pointwise convergence.

The one most people use, and that I'll be using, is the compact-open topology.

These are refinements of each other as so:

- uniform convergence $\supset$ open-compact $\supset$ pointwise convergence.

Thus, if we could find a surjective $s : X \to C(X, I)$ in the uniform convergence topology, it would still be continuous in the compact-open topology. Conversely, if we could show that no such maps exist in the pointwise convergence topology, there would also be no maps in the compact-open one. Unfortunately, the partial results I've found are opposite: I believe I can show that no maps exist for general T_4 X's for uniform convergence, and I have a vague argument that may allow us to construct one in the pointwise convergence topology. Neither of these help here.

Anyway, onwards and upwards!

The steps of the proof

The proof will have lots of technical terminology (with links) but will be short on detailed explanations of this terminology.

It will first assume that X is Fréchet–Urysohn, σ-compact, and a T_4 space. Once the proof is completed, it will then go back and remove the T_4 condition.

Why do it in that order? Because without the T_4 condition, it is a proof by contradiction, assuming a continuous surjection $s : X \to C(X, I)$ to do most of the work. I have nothing against proofs by contradiction, but, if ever anyone wants to refine or extend my proof, I want to make clear which results are really true for X, and which ones only arise via the contradiction.

The proof will proceed by these steps:

- Excluding discrete spaces, and finding compact $Y \subset X$.

- The restriction map $C(X, I) \to C(Y, I)$ is continuous surjective.
- Compact subsets of $C(Y, I)$ have empty interior.
- $C(Y, I)$ is a Baire space.
- There is no surjective continuous map $s : X \to C(X, I)$.
- Removing the T_4 condition.

Excluding discrete spaces, and finding compact $Y \subset X$

If X is discrete, then $C(X, I) = I^X > 2^X$, which has strictly higher cardinality than X, so surjective maps $X \to C(X, I)$ are impossible at the set level.

So put the discrete sets aside. Since X is not exclusively made up of isolated points, it must contain a limit point; call it y_∞. Since X is Fréchet–Urysohn, there exists a sequence $\{y_i\}_{i \in N}$, $y_i \neq y_\infty$, that converges on y_∞.

Then define $Y = \{y_i\}_{i \in N} \cup \{y_\infty\}$; it's clear that it is compact in the subspace topology. Since X is also T_2, Y must be closed in X.

The restriction map $C(X, I) \to C(Y, I)$ is continuous surjective

Any function $f : X \to I$ is a map $f : Y \to I$ by restriction, so there is a map $r : C(X, I) \to C(Y, I)$.

Since X is T_4 and Y is closed, for any continuous function $f : Y \to I$, there exists, by the Tietze extension theorem, a continuous $F : X \to I$ such that $f(y) = F(y)$ for all $y \in Y$. So r is surjective.

We now want to show that r is also continuous.

A subbase for $C(Y, I)$ consists of all $V_Y(K, U) \subset C(Y, I)$ where $K \subset Y$ is compact in Y and $U \subset I$ is open, and $f \in V_Y(K, U)$ iff $f(K) \subset U$.

Note that since Y has the subspace topology in X, these K must be compact in X as well. So define $V_X(K, U)$, consisting of continuous functions F from X to I with $F(K) \subset U$. Then $p^{-1}(V_Y(K, U)) = V_X(K, U)$, and thus the pre-image of the sets of this subbasis is open.

Since it suffices to check continuity on a subbasis, this shows that r itself is continuous.

Compact subsets of $C(Y, I)$ have empty interior

We aim to show, using Ascoli's Theorem, that any compact subset of $C(Y, I)$ has empty interior - thus contains no open sets.

The subbasis for the topology on Y consists of sets of the form $V(K, U)$, where K is compact in Y, U is open in I, and $f \in V(K, U)$ iff $f(K) \subset U$.

The compact subsets of Y are a) finite collections of points y_i, $i < \infty$, b) collections of points that include all points y_i for $i > N$, along with y_∞, and c) y_∞ itself.

Let W be an intersection of finitely many $V(K, U)$. If W is non-empy, it is defined by an $N \in \mathbb{N}$ and a collection of sets $\{U_i\}_{i<N} \cup \{U_{<\infty}\} \cup \{U_\infty\}$ open in I. Then f belongs to W iff:

- $\forall i < N : f(y_i) \in U_i$,
- $\forall i \geq N : f(y_i) \in U_{<\infty}$,
- $f(y_\infty) \in U_\infty$.

Since W is non-empty and f is continuous, we must have $\overline{U_{<\infty}}$ and U_∞ having non-zero intersection. This actually implies that $U_{<\infty}$ and U_∞ have non-zero intersection.

Note that this includes all possible W, as we can always set $U_i = I$ (or $U_{<\infty} = I$ or $U_\infty = I$), thus removing that particular restriction.

We want to show that such an W cannot be equicontinuous at y_∞.

To do so, fix points $u_i \in U_i$ for $i < N$, and $u_\infty, u' \in U_{<\infty} \cap U_\infty$, with $u_\infty \neq u'$

Then consider the functions $f_m \in W$, for $m > N$, defined by:

- $f_m(y_i) = u_i$ for $i < N$,
- $f(y_i) = u'$ for $N \leq i \leq m$,
- $f(y_i) = u_\infty$ for $i > m$.

These f_m are all in W, and continuous (since Y is discrete except at y_∞). If W were equicontinuous then for all $\epsilon > 0$, there would exists an open set U in Y containing y_∞ such that $|f_m(y) - f_m(y_\infty)| < \epsilon$ for all m and all $y \in U$.

Since U is open, there exists a $y_n \in U$, $n \neq \infty$. Set $\epsilon = |u' - u_\infty|/2$; then $|f_n(y_n) - f_n(y_\infty)| = |u' - u_\infty| > \epsilon$.

Since W is not equicontinuous, and Y is both compact and Hausdorff (and hence pre-compact Hausdorff), W cannot be contained in a compact subset of $C(Y, I)$, by Ascoli's Theorem.

Any open set in $C(Y, I)$ consists of unions of sets of the type W, so no (non-empty) open set is contained in compact subset of $C(Y, I)$.

$C(Y, I)$ **is a Baire space**

By the Baire category theorem, if $C(Y, I)$ is a complete metric space, then it is a Baire space.

Since Y is compact, the compact-open and the uniform convergence topology match up on $C(Y, I)$. So $C(Y, I)$ has a uniform norm d compatible with its topology: $d(f, g) = \sup_{y \in Y} |f(y) - g(y)|$.

So $C(Y, I)$ is thus a metric space. Since I is complete, the uniform limit theorem shows $C(Y, I)$ is a complete metric space (and hence a Baire space).

There is no surjective continuous map $s : X \to C(X, I)$

Let $s : X \to C(X, I)$ be any continuous map.

Since X is σ-compact, $X = \bigcup_{i \in N} K_i$ for compact U_i. The set $r \circ s(K_i)$ is compact in $C(Y, I)$, since the continuous image of compact spaces are compact.

Let $L_i = C(Y, I) - r \circ s(K_i)$. Since compact sets in $C(Y, I)$ must have empty interior, the complement of L_i has empty interior. Therefore the L_i are dense in $C(Y, I)$.

The space $r \circ s(X)$ is equal to:

- $\cup_{i \in N}(r \circ s(K_i)) = \cup_{i \in N}(L_i)^c = (\cap_{i \in N} L_i)^c$, where c denotes taking the complement in $C(Y, I)$.

But, since $C(Y, I)$ is a Baire space, $\cap_{i \in N} L_i$ is dense, hence certainly not zero, and so $r \circ s(X) \neq C(Y, I)$.

Since r is surjective $C(X, I) \to C(Y, I)$, the s cannot be surjective onto $C(X, I)$, proving the result.

Removing the T_4 condition

We only used T_4 to prove the properties of the subsequence Y. Without T_4, we can use the continuous surjection $s : X \to C(X, I)$ itself. So now assume that such an s exists, and we shall try and find a contradiction.

The σ-compact property implies X is Lindelöf. Lindelöf is preserved by continuous maps, so because of s, $C(X, I)$ is also Lindelöf.

Because I is T_3, then so is $C(X, I)$. But any Lindelöf regular space is normal (see theorem 14), hence, since $C(X, I)$ is also T_2, then it must be Hausdorff normal, ie T_4.

Now $C(X, I)$ contains at least the constant functions, so must be of uncountable cardinality. Since $X = \bigcup_{i \in N} K_i$ is a countable union of spaces, there exists a K_i such that $s(K_i)$ contains a set W with infinitely many points. Pull these infinitely many points back to K_i, using the axiom of choice to choose a set $V \subset K_i$ such that for each $w \in W$, there is a unique $v \in V$ with $s(v) = w$.

Now, K_i is compact, which means that it is countably compact. Now, Fréchet–Urysohn is hereditary, meaning that it applies to subspaces as well, so K_i is Fréchet–Urysohn, which means that it is sequential. For sequential spaces, countably compact implies sequentially compact (see theorem 10).

Thus we can pick a sequence in V, and, passing to a subsequence if necessary, we can find a sequence $\{y_i\}_{i \in N}$ converging to a point y_∞ that is not equal to any of the y_i. Define Y as before.

The set $s(Y)$ must also be a convergent subsequence.

Let $f \in C(Y, I)$, then $f \circ s^{-1} \in C(s(Y), I)$ (note that s^{-1} is well defined here, as it is a bijection between Y and its image). Because $C(X, I)$ is itself T_4, there exists a function $F \in C(C(X, I), I)$ that is equal to $f \circ s^{-1}$ on $s(Y)$. Then the function $F \circ s$ is equal to f on Y, and is an element of $C(X, I)$. Thus the restriction map $r : C(X, I) \to C(Y, I)$ is still

surjective, and the topology on Y (indeed on any sequence tending to a point) is the same as before. The argument for r being continuous goes through as before.

Then the rest of the proof proceeds as above.

On Abstract Systems

We've all seen those abstract systems that are used for analysis like: Strategy = Ends + Ways + Means, Waterfall Model: Requirements, System Design, Implementation, Integration & Testing, Deployment, Maintenance, SWOT: Strengths, Weaknesses, Opportunities, Threats, ect.

Classes that teach these tend to be boring. Often you'll get three different business models thrown at you with minimal motivation. You won't be told about the environment this model evolved in, nor why the particular elements were included. Sometimes you'll get a concrete example, sometimes not, but if so, it's more likely to be a made up scenario rather than an example from someone's real life experience. It is even rarer that you will receive multiple such examples.

This is largely a result of people's bias towards explicit knowledge. But the explicit knowledge is in most cases simply one of many ways of carving up possibility space. Much more important is the implicit knowledge of how to apply the system and what situations it is helpful in. Strategy = Ends + Ways + Means becomes more applicable when you learn that it became dominate after the Vietnam War when it was felt that the loss was the result of attempting to achieve certain objectives without sufficient resources. So instead of just asking, "What are we trying to achieve?" and "How could we achieve it?", you also wanted to ask, "What resources are needed to achieve it?". Similarly, the Waterfall Model becomes more useful when you are told that many projects have not been as successful as desired because of a failure to consider requirements (ie. using a framework that requires Internet Explorer 10+ when some staff are stuck on Internet Explorer 9). Similarly, people often perform the cost analysis of projects for just the initial coding, without accounting for the maintenance cost. This is a system for making sure that you don't forget the obvious.

It would be even better if I could illustrate these with examples from my own experience, but this should still be sufficient to illustrate my point. I do think that these systems can be useful. But only when communicated in the right way.

This post was written with the support of the EA Hotel

What is narrow value learning?

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

Ambitious value learning aims to achieve superhuman performance by figuring out the underlying latent "values" that humans have, and evaluating new situations according to these values. In other words, it is trying to infer the criteria by which we judge situations to be good. This is particularly hard because in novel situations that humans haven't seen yet, we haven't even developed the criteria by which we would evaluate. (This is one of the reasons why we need to model humans as suboptimal, which causes problems.)

Instead of this, we can use *narrow value learning*, which produces behavior that we want in some narrow domain, without expecting generalization to novel circumstances. The simplest form of this is imitation learning, where the AI system simply tries to imitate the supervisor's behavior. This limits the AI's performance to that of its supervisor. We could also learn from preferences over behavior, which can scale to superhuman performance, since the supervisor can often evaluate whether a particular behavior meets our preferences even if she can't perform it herself. We could also teach our AI systems to perform tasks that we would not want to do ourselves, such as handling hot objects.

Nearly all of the work on preference learning, including most work on inverse reinforcement learning (IRL), is aimed at narrow value learning. IRL is often explicitly stated to be a technique for imitation learning, and early algorithms phrase the problem as *matching* the features in the demonstration, not exceeding them. The few algorithms that try to generalize to different test distributions, such as AIRL, are only aiming for relatively small amounts of generalization.

(Why use IRL instead of behavioral cloning, where you mimic the actions that the demonstrator took? The hope is that IRL gives you a good inductive bias for imitation, allowing you to be more sample efficient and to generalize a little bit.)

You might have noticed that I talk about narrow value learning in terms of actual observed behavior from the AI system, as opposed to any sort of "preferences" or "values" that are inferred. This is because I want to include approaches like imitation learning, or meta learning for quick task identification and performance. These approaches can produce behavior that we want without having an explicit representation of "preferences". In practice any method that scales to human intelligence is going to have to infer preferences, though perhaps implicitly.

Since any instance of narrow value learning is defined with respect to some domain or input distribution on which it gives sensible results, we can rank them according to how general this input distribution is. An algorithm that figures out what food I like to eat is very domain-specific, whereas one that determines my life goals and successfully helps me achieve them in both the long and short term is very general. When the input distribution is "all possible inputs", we have a system that has good behavior everywhere, reminiscent of ambitious value learning.

(Annoyingly, I defined ambitious value learning to be about the *definition* of optimal behavior, such as an inferred utility function, while narrow value learning is about the observed behavior. So really the most general version of narrow value learning is

equivalent to “ambitious value learning plus some method of actually obtaining the defined behavior in practice, such as by using deep RL”.)

A general framework for evaluating aging research. Part 1: reasoning with Longevity Escape Velocity

Summary

To this day there is a lack of systematic research to evaluate a cause area with immense potential: aging research. This is the first of a series of posts in which I'll try to begin research to address this gap. The points made in this post are about how to evaluate impact using the concept of Longevity Escape Velocity. Bringing the date of Longevity Escape velocity closer by one year would save 36,500,000 lives of 1000 QALYs, using a conservative estimate. Other sources of impact that arise from the same concept include: increasing the probability of Longevity Escape Velocity, making Longevity Escape Velocity spread faster, and making a new future portion of the population reach Longevity Escape Velocity by increasing its life expectancy. Aging research could also positively impact the cost-effectiveness of other interventions by increasing the probability that Longevity Escape Velocity will be attained in the recipients' lifetimes. I will also discuss why the probability of Longevity Escape Velocity is substantial and why QALYs should be the measure of impact, and I'll give mathematical proofs that the adoption speed of the technologies that arise from research doesn't impact cost-effectiveness analyses.

The need of a theoretical foundation to evaluate aging research

I think one important approach to research in Effective Altruism is to try to lay theoretical foundations and put together tools for helping to evaluate a specific cause area that can be generalised to any intervention inside that cause area. Such work is often not possible because of lack of time and expertise, making it preferable, sometimes, to scout specific promising interventions or refine existing research.

One cause area that absolutely needs this kind of more systematic groundwork is aging research. The current EA research about aging is lacking in number and in what I think are crucial considerations, even though informal discussion with members of the community reveals that many people regard it as potentially promising. The expertise required to make such an analysis possible is rare to find. It requires people with a strong quantitative background who are also interested not only in biology but in the biology of aging in particular, and they must be accustomed to predicting the future of scientific research and making cost-effectiveness evaluations. I observed that the Effective Altruism community seems to have plenty of people with a background in philosophy, economics, social sciences or computer science, but people with a strong background in biology, or at least a strong interest in it, are scarce. This makes it even harder to find people willing to do the work of evaluating the cause area of aging research.

For these reasons, and since I'm very familiar with the topic and I think I have important things to say about it, I am willing to try to lay as much groundwork as possible, at least until I think I'm needed.

My long-term hope is that the groundwork I will lay will be good enough for a more formal discussion about this topic within Effective Altruism, both for evaluating specific interventions inside this cause area and for evaluating the cause area as a whole. I will write about what I think are original points and put together all the existing tools that could help both Effective Altruism organisations and organisations within the cause area of aging to make better decisions.

I have chosen to split the analysis into multiple posts so that I can receive and incorporate feedback during the process and thereby modify my work and its planning along the way. Organising the work in this way will also make the whole thing easier to read.

I'm doing this alone and in my free time, between university and other activities, so the posts will probably come out a few weeks or even months apart.

Although I hope that the bottom line of my arguments is strong, there will probably be many mistakes and many corrections to make. I encourage you to comment, give feedback and to contribute new ideas, especially if you have a consideration about something that I didn't address that would substantially impact the result of a potential cost-effectiveness analysis. Along the way I will probably need to collaborate with other people and coauthor some posts, since my knowledge probably has gaps and needs to be complemented. Nonetheless I will try to learn what I currently don't know along the way.

What will this series of posts be about?

This is the first of a series of posts in which I'll explore different ways of reasoning about the potential cost-effectiveness of aging research. Each post will focus on one or more considerations. In the last post I would like to wrap everything up with a comprehensive framework useful for evaluating the cost-effectiveness of any given avenue of scientific research into the aging processes and how to treat their various facets. The points made will also provide an idea of the potential of the cause area as a whole.

An initial major focus will be on the scope of the problem and on moral considerations that could affect it. Neglectedness and tractability will be given space in later posts, in which I will try to lay out useful methods and heuristics to evaluate them in this cause area.

After this work, I would like to discover the best funding opportunities within this area and compare my conclusions with other past efforts within Effective Altruism that have been made to evaluate aging research.

Points made in this first post:

- Longevity Escape Velocity (LEV) is the minimum rate of medical progress such that individual life expectancy is raised by at least one year per year if medical interventions are used.
- Reasoning with Longevity Escape Velocity substantially changes cost-effectiveness analyses.
- A conservative estimate for life expectancy after Longevity Escape Velocity is 1000 years, although it’s still not a lower bound.
- In order to account for making Longevity Escape Velocity arrive more quickly in cost-effectiveness analyses (CEAs) the relevant variables are deaths by aging per year, life expectancy after LEV and Expected number of years LEV is made closer by. This without accounting for moral weights and other possible discounts.
- The probability of Longevity Escape Velocity is substantial.
- Another factor potentially greatly influencing impact is the life expectancy increase resulting from research projects or health interventions. If the project is not likely to be funded in the future or subsumed by other research, the recipients of the intervention who would have died near LEV get saved.
- QALYs should be the measure of impact, as one life saved counts more than 30-80 QALYs in this cause area.
- Mathematical proofs that cost-effectiveness analyses aren't influenced by how Longevity Escape Velocity, or any technology that arises from a given financed research project, spreads to the whole population
- Making LEV spread faster is another impact consideration that is pertinent to projects potentially leading to policy change or public awareness.
- Certain research projects could also have the effect of increasing the probability of Longevity Escape Velocity, potentially influencing cost-effectiveness analyses substantially.
- Aging research can boost the impact of other altruistic interventions by increasing the probability of LEV happening in the recipients' lifetimes.

The next posts in the series will probably be about:

- A better lower bound for the life expectancy after Longevity Escape Velocity, and how this affects the probability of LEV.
- The longevity dividend.

- The value of information (depending on if I need to include considerations specific enough to aging research).
- Moral weights.
- If old people are replaceable in a utilitarian moral framework.
- What is neglected and what is tractable in the cause area of aging research.
- Putting the framework together.

Longevity Escape Velocity: what it is

Longevity Escape Velocity (LEV) is the minimum rate of medical progress such that individual life expectancy is raised by at least one year per year if medical interventions are used. This does not refer to life expectancy at birth; it refers to life expectancy calculated from a person's statistical risk of dying at any given time. This is equivalent to saying that a person's expected future lifetime remains constant despite the passing years.

It's possible, given sufficient ongoing improvement of medicine and its democratisation, that nearly everyone on the planet, at a certain date in the future, will benefit from therapies that allow Longevity Escape Velocity to be attained, at least until aging is eradicated completely. Then, other factors will influence risk of death, and expected future lifetime could start falling again each passing year if risk of death flattens or doesn’t continue to fall fast enough.

How likely that Longevity Escape Velocity is to become a reality in the future depends on a number of factors, which will be explored later in this post.

Reasoning with Longevity Escape Velocity substantially changes cost-effectiveness analyses

If a given intervention "saves a life", this usually means that it averts 30 to 80 Disability-Adjusted Life Years (DALYs). In order to evaluate the impact of aging research, one could be tempted to try to estimate how many end-of-life DALYs that a possible intervention resulting from the research could save and adjust the number using the probability of success of the research.

This is the line of reasoning that OpenPhilanthropy's medium investigation on aging uses, although without making any explicitly quantitative argument. This is part of the impact, and it has to be factored in, but it doesn't consider where the largest impact of aging research is: making the date of Longevity Escape Velocity come closer. This would have the effect of saving many lives from death due to age-related decline and disease, but here, "a life" means, more or less, 1000 Quality-Adjusted Life Years (QALYs). This figure is derived as follows:

Life expectancy after LEV:

In actuarial science, the expected future lifetime of an individual at age x is denoted with e_x. It can be seen as the expected value of the random variable K(x), also called "Curtate Future Lifetime", which is defined as $K(x) := \lfloor T(x) \rfloor$, where T(x) maps to the amount of additional time that an individual of age x is projected to live. For our purposes, we can use the discrete random variable K(x) instead of directly T(x). Thus, with ${}_kp_x$ being the probability of surviving between age x and age x + k:

$$e_x = E[K(x)] = \sum_{k=0}^{\infty} k \cdot P\,(K(x) = k) = \sum_{k=0}^{\infty} k \cdot ({}_kp_x - {}_{k+1}p_x) =$$

$$= \sum_{k=0}^{\infty} k \cdot {}_kp_x - \sum_{k=0}^{\infty} k \cdot {}_{k+1}p_x = \sum_{k=1}^{\infty} {}_kp_x$$

If p(x) is the probability of dying between year x and year x + 1, then:

$$e_x = \sum_{k=1}^{\infty} {}_kp_x = \sum_{k=1}^{\infty} \prod_{i=1}^{k} (1 - p(x + i - 1))$$

When the whole population benefits from LEV, the risk of death will fall for everyone. By definition, it will fall at a rate such that the expected future lifetime of any given individual will remain constant until aging gets eradicated completely. So, in order to make the most conservative estimate about life expectancy after LEV, we need to find the minimum rate of decrease of p(x) such that this condition holds. The answer to this doesn't seem easy, so I'll find this lower bound in another post.

For now, I'll use a constant risk of death to calculate the life expectancy of individuals after aging is eradicated completely and risk of death has presumably stopped falling. While this method doesn't yield a lower bound, since it leaves out from the calculation the risk of death when it's decreasing, it can be made conservative using a relatively high risk of death. I'll use p(x) = 1/1000, which is more or less the current risk of death of someone between 20 and 30 years old. It is conservative because it doesn't account for future improvements in medicine and general safety outside of aging research. I also don't expect the lower bound to be much smaller. Therefore,

$$e_x = \sum_{k=1}^{\infty} {}_kp_x = \sum_{k=1}^{\infty} (1 - 1/1000)^k = 999 \approx 1000$$

Since we are talking about life expectancy in a world without aging, 1000 years of life expectancy should amount more or less to 1000 QALYs.

Accounting for making LEV come closer in CEAs

Any given aging research project, if successful, could have the effect of making the date at which most people will reach Longevity Escape Velocity come closer by a certain amount of time. We can estimate the expected QALYs gained because of such an effect. We have established that the average lifespan of a person who reached LEV will be around 1000 years, mostly without disability, and somewhat less if we use a lower bound. The number of QALYs saved are then calculated by multiplying 1000 by the number of people who would otherwise have died of aging if LEV wasn't moved closer. Currently, around 100,000 people per day (36,500,000 people per year) die due to age-related decline and diseases, although this figure will be larger when LEV arrives due to population growth.

So, in order to calculate an almost lower bound for how many expected QALYs that a certain research project would save by making LEV come closer, you simply multiply these values:

- 1000 QALYs
- 36,500,000 deaths/year
- Expected number of years LEV is made closer by

This is true for a crude estimate, without accounting for moral weights and potential discount rates.

It is important to stress the fact that none of these variables depend on how soon LEV will arrive, so we can totally ignore this kind of discussion, even if it is a highly debated topic outside the

setting of cost-effectiveness evaluations.

The first two variables have been already discussed. Then, we need to examine the third one, which depends on many factors, such as:

- How promising the project being examined is, for different meanings of the word "promising". For example, it may have direct translational value into effective therapies targeting aging processes or hallmarks, or it could have an effect of speeding up the field, such as by providing new tools, by enabling political entities to aid in achieving LEV sooner, or by enabling a new line of research to start sooner or become widespread more rapidly.
- The neglectedness of the project would make the figure larger.
- If there are other projects that could subsume the effect of the examined project, some of which would universally subsume all potential projects. These include technologies outside the field of aging research that are potentially very disrupting and sudden, such as artificial general intelligence.
- The number of years necessary for another group to step in and do the same project.
- The probability of catastrophic events: existential risks or events catastrophic enough to make the information acquired by the project lost or useless.
- Lastly, the probability that LEV will happen in the first place also has a role in estimating this variable. This is because we can model the number of years that LEV is brought closer as the expected value of the random variable that maps to various numbers of years, among which is zero. The probability that the years LEV is brought closer by is zero, in turn, depends not only on the specifically examined project but also on the probability that LEV will not happen. That's why it's useful to outline how to reason about the probability of LEV.

Probability of LEV

If we had the minimum rate of decrease of risk of death such that LEV would happen, then the probability of LEV happening is the probability that the risk of death would fall at that rate or faster, and so the probability largely depends on that rate and on how fast medical research will be.

For now, we can reason about the problem by dividing the situations in which LEV will not happen in at least two scenarios:

- Very slow research scenario: In this scenario, each new therapy is developed in the span of an entire generation and contributes only a few more years of healthy life. This slow rate of progress is maintained more or less constantly within the span of a few centuries. Example: every 30 years or so, only one new therapy grants 5 more years of healthy life for the general population. If progress is this slow, LEV will never be reached. Negligible senescence will eventually be met after a few centuries, and generations will have progressively longer lifespans without anyone suddenly making very large jumps in life expectancy. New therapies may rely on previous ones for having substantial effects, forcing new treatments to come sequentially and not in parallel. It's also possible that they could theoretically be developed in parallel, but an incredibly inefficient research community develops them sequentially. This scenario seems somewhat unlikely. This tells us that reaching the minimum rate of decrease of risk of death shouldn't be too difficult.
- Dire roadblocks scenario: This is the scenario in which there are roadblocks so dire that aging research is stalled for enough time that the recipients of previous interventions die. This doesn't necessary prevent LEV all the time; these kind of roadblocks must be enough in number to effectively make the average decrease of risk of death the same as the one of the very slow scenario until aging is cured completely.

The scenarios in which LEV will happen, instead, are the ones in which the risk of death falls fast enough, which means that new therapies would be developed sufficiently close together. This would be brought about through steady progress in medicine or relatively large jumps in life expectancy that enable previous recipients of therapies to extend their lives by another large amount of time. We can imagine how such scenarios could unfold:

- Today's therapies or future therapies appear to be somewhat effective on humans or very effective on mice. This increases public focus on translational aging research, which, in turn,

results in a multiplication of resources dedicated to it. It's argued that that this first "proof of concept" required to convince the world is Robust Mouse Rejuvenation, which would double the remaining life expectancy of elderly mice, as demonstrated and then replicated in rigorous laboratory studies. A multiplication of resources for the field should result in therapies following the first proofs of concept. After this, the rate of therapy development and improvement will increase exponentially following the initial success of therapies. The history of technology is full of examples of this feedback loop, in which successive improvements are faster than the development of the proof of concept, a prominent one being flight.

- Without invoking a large public interest, LEV could also be caused by combinations of different treatments coming in waves and by the improvement of treatments over time. This would mean sudden jumps in life expectancy that would buy enough time for other treatments to be developed. A sudden future enlightenment about the nature of aging could be also possible, or the first therapies could also have the effect of slowing down the accumulation of other damages, other than doing the job of addressing their specific targets. This would happen by breaking negative feedback loops of damages or processes. LEV could also happen in a sudden way if effective delivery methods are developed after many proof-of-concept therapies have been demonstrated, for example, in vitro.
- The two scenarios above sound somewhat optimistic, but they might not be needed at all. The research could unfold silently but surely and the risk of death could still fall fast enough to ensure LEV. This would happen if the current situation of very slow improvement is overcome and there isn't a large number of new dire roadblocks ahead.

Given these scenarios, can we have a preliminary idea, without knowing how fast the risk of death needs to fall, of how likely LEV is? There are, at least, probably some relevant points to make regarding the current best guesses about aging and the present state of research.

It's difficult to predict major future roadblocks, but at least it seems that the "very slow research" scenario is proving less and less likely. This doesn't mean that we already have effective therapies against aging, or that the pace of science is optimal. But how research is distributed and the theories about what aging is make believable the possibility of therapies being developed closer together, thereby enabling a high-enough rate of decrease of overall risk of death.

The current best guess about how to tackle aging rests on a milestone paper from 2013: The Hallmarks of Aging. The paper has citations in the thousands and counting, and researchers are using it as a framework to orient and justify their own research. It proposes various categories of dysfunction. Every category, or almost every category, should be addressed periodically in order to maintain a youthful state of health. Reversing one hallmark would mean restoring an internal state of the body that is typical of a youthful body. It could also prove true that it will not be necessary to address every hallmark, due to the possible cause-effect relationships between each of them.

What does this say about how close together therapies will come? It says a lot: a paper like The Hallmarks of Aging means that the field already has an idea of what combination of foreseeable therapies will bring major gains in health and, in turn, life expectancy. This is because this theoretical categorisation constitutes what needs to be addressed.

It also implies that it enables thinking about rejuvenation, not only "slowing down" aging. This is because the dysfunctions described are exactly what is "wrong" with an old body, and not how those dysfunctions arise, so getting rid of those kind of dysfunctions means rejuvenation.

It's a "downstream" view of aging that decomposes the problem and leaves out what is unnecessary to know in order to intervene, increasing the tractability of the problem. We don't need to know how the Hallmarks arise in order to develop therapies that address them. One added benefit is that the hallmarks influence each other in negative feedback loops; reversing one slows down the progress of many others.

Theoretically, interventions aiming at reversing all of the hallmarks of aging could be developed in parallel, and, in fact, they currently are (although not optimally so). Interventions to ameliorate each one of the Hallmarks, at least in specific parts of the body, are underway. You can follow the progress of each research targeting each hallmark by using the Rejuvenation Roadmap made by the Life Extension Advocacy Foundation. This map tracks the progress of research projects that

ameliorate each hallmark and provides links with explanations of each project; it also contains citations to the relevant papers.

As you can see, there are some hallmarks, such as mitochondrial dysfunction and loss of proteostasis, which are in the very early stages of research: the furthest they have reached, so far, is the preclinical stage. Research on how to ameliorate mitochondrial dysfunction, in particular, is in such an early stage of research that it is only pursued by nonprofits and academia, but it needs to be addressed in the wider scheme of therapies that will be needed in order to address all of the dysfunction arising from aging.

There are other hallmarks, such as cellular senescence and stem cell exhaustion, which are in fairly advanced stages of research (phase 1 and phase 2 trials), and research on them is pursued by well-funded, for-profit companies, such as Unity Biotechnology.

The fact that all of these lines of research are pursued in parallel is important. It means that at an unspecified time in the future, near or far, lines of research could come together in a relatively short period of time. The fact that right now, many interventions are being researched on specific diseases (e.g. Unity Biotechnology's trial is for arthritis) does not negate the previous point: treatments that are being researched using the Hallmarks framework, even though they are being tested for specific conditions, are relevant for therapies that treat a wide range of diseases. Parallel development makes it more likely that therapies will come in waves, with each therapy being released shortly after another.

There are also other approaches in aging research, such as targeting aging in a more upstream fashion, with less ambitious interventions that target metabolic pathways. One example is metformin, although I don't think that, right now, science is advanced enough for research on specific medical interventions using this approach to substantially make the date of LEV come closer or substantially impact its probability. These kinds of research projects, nonetheless, could have the effect of buying some time for an additional slice of the population to reach LEV. This brings us to another way of accounting impact in this cause area.

Accounting for making an additional slice of the population reach LEV

Another factor potentially greatly influencing impact is the life expectancy increase resulting from research projects or health interventions. If the project is not likely to be funded in the future or subsumed by other research, the recipients of the intervention who would have died near LEV get saved. I think the health interventions or projects for which this factor is relevant are very few or maybe even non-existent. This consideration influenced the impact measure I used in my previous analysis on the TAME trial, but in retrospect I think I overestimated the probability that the health benefit of metformin will not be subsumed by other research.

In order to account for this, the relevant factors to multiply are:

- Life expectancy after LEV.
- Recipients of the interventions who would have died just before LEV if their life expectancy wasn’t extended by the intervention.
- Probability that the project will not be funded by someone else, or is subsumed by other research.

QALYs should be the measure of impact

Due to the possibility of LEV, expected QALYs should be the measure of impact of aging research. Lives saved lose their original meaning, unless 1 life of 1000 QALYs is counted as multiple lives of 30-80QALYs. Exactly how many also depends on how moral weights are chosen. In my previous analysis about the cost-effectiveness of the TAME trial, I made the mistake of measuring impact in lives saved instead of directly in QALYs, without considering the fact that a life saved in that context amounted to 1000 or more QALYs and actually counted as multiple lives saved. In that

analysis, I also didn't account for DALYs averted at the end of life and every other factor that influences impact, which I will discuss in future posts.

How LEV spreads will have no impact on CEAs

A concern sometimes comes up when I present LEV-based reasoning: how do we account for the fact that LEV will probably spread to the whole population over a large period of time (e.g. following the sigmoid technology adoption curve)? This consideration has no effect on the final estimate of cost-effectiveness, and making the date of LEV closer by any given period of time prevents exactly the number of deaths by aging occurring during that period of time, regardless of how LEV spreads. Let's first see a simple example where this holds and then prove it mathematically in the general case. I came up with two proofs, each one of which is sufficient alone.

These same arguments and proofs work for any other outcome of a given technology. How a certain technology (health-related or not) will spread doesn't influence the cost-effectiveness of financing the research leading to it. I may include the generalised version, which trivially follows from this one, in a separate post.

Keep in in mind that this does not mean that making LEV spread faster doesn’t impact CEAs. In fact, this is a potential impact factor that I will discuss. This result means that the impact of making the date of LEV come closer isn't influenced by how LEV spreads.

I will use deaths prevented in the example and in the proofs, but a generic measure of impact yields the same result. Using QALYs is not necessary in this case.

Example

Let's consider two specific scenarios as an example: in the first scenario, LEV spreads to the whole population instantly, and in the second, it spreads over four years.

First scenario: A particular piece of research makes LEV come closer by one year. Since LEV spreads instantly over the whole population, it's easy to see that the resulting deaths prevented are exactly the deaths by aging occurring during one year: more or less 100k.

Second scenario: A particular piece of research makes LEV come closer by one year, but LEV spreads over the world during a period of four years. In the first year, 1/4 of the population reaches LEV; in the second year, 1/2; in the third, 4/5; and in the fourth, 5/5. If we shift this gradual transition by one year, then in the first year, we prevent, on the margin (deaths that would have occurred if we didn't move the date), 1/4-0 = 1/4 of the deaths of aging occurring during that year. In the second year, we prevent 1/2-1/4 = 1/4 of the deaths by aging that occur during that year. In the third year, we prevent 4/5 - 1/2 = 3/10 of the deaths by aging occurring during that year. In the fourth year, we prevent 5/5 - 4/5 = 1/5 of the deaths by aging occurring during that year. So, in total, by shifting the date of LEV by one year, we prevented: 1/4 + 1/4 + 3/10 + 1/5 = 1. That is, we prevented the deaths by aging occurring during one year: more or less 100k.

As you can see, the number of deaths prevented in the two scenarios is the same: the number of deaths by aging occurring during one year. LEV is moved closer by one year in both scenarios, but it spreads differently.

Now, I'll prove, more generally, that making LEV closer by any given period of time prevents exactly the number of deaths by aging occurring during that period of time, regardless of how LEV spreads.

Proof 1:

n = the number of years needed for therapies to spread to the whole population.

y =the year in which the therapies leading to LEV begin spreading.

d = number of deaths caused by aging each year (d could be the number of deaths by aging occurring in any arbitrary unit of time; the proof remains the same).

$f : \{y, \ldots, y + (n - 1)\} \to [0, d]$ expresses how many deaths from aging are prevented in a given year during the time that therapies are spreading. How exactly f is defined depends on how the therapies spread (e.g. exponentially or linearly), but we know that $f(y + n - 1) = d$ and that $f(y) = 0$.

If LEV spreads to the whole population all at once, then $n = 1$ and $f : \{y\} \to \{d\}$. In this case if the date of LEV is moved closer by 1 year, then the resulting new function $f^{(1)}$, has $y - 1$as the only member of its domain, also mapping to d. So the deaths prevented on the margin by making the date of LEV closer by one year are exactly d.

We want to prove for all values of n and y that if the date of LEV is moved closer by one year but the therapies do not spread to the whole population all at once, the number of deaths prevented on the margin still amounts to d.

Let $f^{(1)}$ be the function that expresses deaths by aging prevented each year after making LEV come closer by one year and f the (already defined) function that expresses deaths by aging prevented each year without LEV being moved closer. Therefore $f^{(1)}$ has the following properties:

$f^{(1)} : \{y - 1, \ldots, y + n - 2\} \to [0, d]$

$f^{(1)}(y - 1) = f(y), f^{(1)}(y) = f(y + 1), \ldots, f^{(1)}(y + n - 2) = f(y + n - 1)$, this holds under the very solid assumption that making the date of LEV closer only shifts f: it doesn't change how it is defined, but only subtracts 1 to all the members of its domain.

Then, the deaths by aging prevented each year on the margin if we make LEV come closer by one year are:

$f^{(1)}(y - 1) + [f^{(1)}(y) - f^{(1)}(y - 1)] + \ldots + f^{(1)}(y + n - 3) + [f^{(1)}(y + n - 2) - f^{(1)}(y + n - 3)] =$

$= [f^{(1)}(y - 1) - f^{(1)}(y - 1)] + [f^{(1)}(y) - f^{(1)}(y)] + \ldots + [f^{(1)}(y + n - 3) - f^{(1)}(y + n - 3)] + f^{(1)}(y + n - 2) =$

$= f^{(1)}(y + n - 2) = f(y + n - 1) = d$

■

Note that the same exact proof works if the date of LEV is moved closer by more or less than one year: It is sufficient to let d be the deaths by aging prevented in an arbitrary unit of time. Another proof, with f having a continuous domain, involves manipulating integrals. Here it is:

Proof 2:

Let f(t) : R → R be the function that associates time with deaths by aging prevented at that time.

Then, the total number of deaths prevented in a given time interval $[t_0, t_1]$ is $\int_{t_0}^{t_1} f(t)\,dt$. The number of deaths averted on the margin if we make the date of LEV come closer by the time interval Δt is:

$$\int_{t_0}^{t_1} f(t + \Delta t)\,dt - \int_{t_0}^{t_1} f(t)\,dt$$

Let's divide the interval $[t_1, t_{n+1}]$in n smaller periods of time of length Δt (the period of time LEV is moved closer by). Let's call those subintervals $\Delta t_i = [t_i, t_{i+1}]$. Then the above integral can be rewritten as a sum of integrals over the smaller intervals.

$$\sum_{i=1}^{n} \left(\int_{t_i}^{t_{i+1}} f(t + \Delta t)\,dt - \int_{t_i}^{t_{i+1}} f(t)\,dt\right)$$

But since it's true that:

$$\int_{t_i}^{t_{i+1}} f(t + \Delta t)\,dt = \int_{t_{i+1}}^{t_{i+2}} f(t)\,dt$$

Then, the terms of the sum simplify with each other and we have:

$$\sum_{i=1}^{n} \left(\int_{t_i}^{t_{i+1}} f(t + \Delta t)\,dt - \int_{t_i}^{t_{i+1}} f(t)\,dt\right) = \int_{t_n}^{t_{n+1}} f(t + \Delta t)\,dt - \int_{t_1}^{t_2} f(t)dt$$

Notice that if t_1 happens one unit of time before LEV begins spreading and $t_{n+1}$is the time at which LEV has reached the whole population, then $\int_{t_1}^{t_2} f(t)dt = 0$and $\int_{t_n}^{t_{n+1}} f(t + \Delta t)\,dt$is exactly the number of deaths by aging that would have occurred in the time interval Δt; this is exactly the number of deaths by aging prevented if LEV was moved closer by Δt and the therapies spread instantly. This proves that the number of deaths by aging prevented on the margin by moving the date of LEV closer by Δt is always equal to the number of deaths by aging occurring during Δt, regardless of how the therapies spread over the world.

■

Accounting for making LEV spread faster

As anticipated, another potential source of impact to consider is if a certain project, for example an advocacy-related or policy change, can change how fast people get access to treatments. This would make LEV spread faster, and, in turn, save the people who otherwise wouldn’t reach it.

In order to evaluate this, we need to come up with an estimate of how the future adoption curve will look like. This could possibly be achieved by looking at the way that current health treatments

spread, and then evaluate how many people, and, in turn, QALYs, get saved by a change in the curve resulting from the project.

This consideration will probably be given more space in a separate post investigating current adoption curves for health-related technologies and the impact of advocacy and inducing policy change in this area.

Accounting for increasing the probability of LEV

Another consideration that may be taken into account to evaluate impact is how much a given research project increases the probability of LEV. This doesn't mean increasing the probability of aging getting eradicated completely; it means increasing the probability of research being fast enough to ensure LEV and not a scenario in which research is so slow, or roadblocks so dire, that aging eventually gets eradicated but no one experiences LEV in the meantime.

This probably depends much on the project in question, but it is also possible that, in general, the impact of this consideration could be small, unless we assume a really inefficient research community, or we are analysing a specific research project that is highly neglected and has the potential effect of removing a major roadblock or unlocking further progress, thus speeding up the general pace of research and making the risk of death for future recipients of interventions fall faster. This could be true for research on new scientific tools or research done in a particularly original way that shows a new approach to problems that wasn't seen before.

In case this consideration has to be taken into account and we need to calculate how many QALYs that increasing this probability would save, we should come up with a distribution of probabilities (with the sum of the probabilities = 1 - the probability of LEV not happening) about how fast the risk of death would fall after LEV. Each outcome yields a different number of QALYs saved by LEV. Then, we should calculate the expected value in QALYs of the distribution with an increased total probability of LEV and the expected value in QALYs of the distribution without an increased total probability of LEV, then determine the difference between the two results.

Aging research boosts the impact of other altruistic interventions

It should be noted that another effect of aging research is to increase the chance for people saved by other interventions to reach LEV. If a life in Africa is saved thanks to insecticidal nets, then the expected QALYs saved will be more or less the person's expected remaining life plus his/her life expectancy after LEV in QALYs (more or less 1000 as we have seen) multiplied by the probability that person has to achieve LEV during the rest of his/her life.

The probability of an individual reaching LEV depends on:

- The probability of dying of causes not related to aging.
- The probability of LEV arriving in the lifetime of the recipient of the intervention.

The first depends on the recipients of the intervention, but even in the worst cases, it should be a pretty high number, considering that even in Africa, the lowest average life expectancy for children born in 2018 is 57 years. The second probability depends on how likely LEV is to appear in any given year. In order to determine this, a very thorough and detailed analysis is needed, and so I will probably tackle this problem in another post.

This is a crosspost from my post in the Effective Altruism Forum.

Failures of UDT-AIXI, Part 1: Improper Randomizing

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

While attempting to construct a UDT-like variant of AIXI, the algorithm I wrote down turned out to have a bunch of problems with it that were blocking attempts to prove that it had various nice properties. This will be the first post in a sequence about the various nonobvious things that go wrong when you try to make UDT-AIXI. (Check back here for more links)

Notation and Setup:

To begin with, the basic setting was one where the Turing machines in the environment had oracle access to the policy. In the first attempt, where O was the set of observations, and O^* was the set of all finite strings of observations, policies had the type signature $\pi : O^* \to \Delta A$, and environments had the type signature $O^* \times \pi \to \Delta O$, and functioned by having the option of making oracle calls to the policy with various observation strings to see what action was returned. If the policy randomizes on that input, then the action is simply drawn from the distribution $\pi(o\!\!\rightarrow)$.

Note that we're just using O^* instead of the AIXI input type of $(O \times A)^*$, this is because the actions that the agent sees on the input tape can just be seen as a sort of observation, like any other, and this generalization permits looking at cases where the agent may not have an internal sense of which action it took, or cases where the agent gets to peek at what it did in hypothetical situations that didn't happen.

Policy Search and Nash Equilibria:

Glossing over issues of nonhalting environments, we can reframe the UDT search for the best policy as the result of an infinite game, where each player is indexed with an observation string $o\!\!\rightarrow \in O^*$, and responds with a (probability distribution over) move(s) in A. Each cell corresponds to a deterministic policy of type $O^* \to A$, and the utility in that cell is the expected reward over the standard mixture of environments, when the environments are making oracle calls to that selected policy. All players conveniently have the same utility function, which makes things easier. The optimal deterministic policy is a Nash equilibrium, because any deviation from any of the players would result in an equal or lower expected utility.

This rephrases problems of selecting a policy, to an equilibrium selection problem, because even when all players have the same utility function, there can be

inadequate equilibria. Consider a multiplayer game of stag hunt where rabbits are shared equally among the group, but it's still better for everyone to team up and capture the stag than for everyone to catch a rabbit. The all-rabbit equilibrium is a Nash equilibrium that isn't the best ash equilibrium. This case seems simple to resolve, but there are much more difficult examples, such as a MAX-3-SAT game where each hypothetical version of the agent in Omega's head has incompatible observation strings, some short, some very long, and they are each responsible for fixing the value of one variable.

Problem 1: Improper Randomization

Setting the Vingean and game theory issues to one side, there's another thing that's going wrong here. *The stated setup of how the policy interacts with the environment is incompatible with this Nash equilibrium view of UDT*.

As a simple example of what's going wrong, consider a single-turn game against an environment that makes two queries to what the policy does, and returns a reward of 1 if they are the same, and 0 if they are different. In standard game theory, randomizing between two actions with identical payoff should have the same expected payoff. But here there's a penalty for randomizing! And because each call to the policy returns an action sampled from the distribution $\pi(e\twoheadrightarrow)$, identical oracle calls to the policy may return different outputs. The proper way to view this to make it compatible with the Nash equilibrium view is to sample the action *first*, and *then* plug it into the open slots in the environment. You only randomize once.

This isn't quite as trivial as it looks, because Reflective Oracles have this "intrinsic randomness" thing going on, two identical oracle calls to the same agent may return different outputs. In Abram's words, reflective oracles are annoyingly realist about probability, they act as if a agent can randomize in such a way as to take different actions in the same information state. The proper way to do this is to sample a deterministic policy first, and then plug it into the slots in the environment. When thinking about a situation where an agent (possibly yourself) is randomizing, you may not know what action it will take, but you should think that the agent will take the same action in an identical state.

The obvious hack is to just cache the result every time the environment makes an oracle call, because if you compute it up-front, the policy sample will probably take an infinite amount of information to specify. But if you apply this hack and cache what was said about the policy so far, you run into another issue with non-halting environments. Because Reflective Oracle-AIXI calls an oracle on an environment to get the next bit, instead of just running the environment (maybe the environment doesn't halt, and you still need to get the next bit somehow), you don't get the policy sample.

If the partial policy sample used by the environment on turn t takes less than f(t) (where f is computable) bits to encode, you can use the oracle to recover it and use that information on the next turn, but again this fails in full generality. The problem arises when you've got an environment that, on some turn, with some positive probability, doesn't halt and makes infinitely many oracle calls. Then there's no way (that I know of, yet) to guarantee that the behavior of the policy on future turns is consistent with what it did on that turn.

Problem 2: Can't Build Desired Oracle

What if we then attempt to get a new notion of oracle where we have a probability distribution over samples? That might have promise, because only a single sample would be used. For instance, the type of a sample would be $M \to \{0,1\}$, and we'd be looking for a fixed-point in $\Delta(M \to \{0,1\})$, ie, a distribution over samples s.t. the probability of drawing a sample that maps M to 0 is the same as the probability of M outputting 0 when it selects a sample from the given distribution.

Then it doesn't straightforwardly follow from the Kakutani fixed-point-theorem, because one of the restrictions is that the set of interest is a nonempty, **compact**, convex subset of a locally convex Hausdorff space. Fixing some bijection between $M \to \{0,1\}$ and $[0,1]$ so we're looking for a fixed-point in $\Delta([0,1])$, we run into the issue that this space isn't compact under any of the natural topologies, and even restricting to computable functions from $M \to \{0,1\}$ (and bijecting that with N), ΔN isn't compact under any of the natural topologies either.

Due to this issue, the earlier post about oracles that result in correlated-equilibria, is incorrect, because the space of interest suffers from this lack of compactness, and the topological preconditions for Kakutani applying weren't checked. The result still holds up for finitely many players with finitely many moves, because we don't need a distribution over N for that, only a distribution over finitely many integers, so the fixed-point theorem goes through in that case.

Optimizing for Stories (vs Optimizing Reality)

Epistemic status: highly confident in the basic distinct, though it's not at all profound. Further details, models, and advice are somewhat speculative despite being drawn from varying amounts of observation. Essay a little rushed since otherwise I'd be unlikely to publish.

When setting out on a venture, one faces some choices.

- **Optimize reality:** How much do you optimize for the success of your stated goal? I define goal to be *some desired state of reality such that success or failure is assessed with respect to reality.*
- **Optimize stories:** How much do you optimize for the appearance of success at your stated goal, i.e. stories? I define a story as *a collection of facts about a reality asserted to be true and relevant, usually presented to others but sometimes only to oneself.*

Ideally, optimizing for one would be the same as optimizing for another. Very often, however, optimizing one is not the same as optimizing for other. What's worse, the two ends compete for the same set of limited resources.

A concrete example might be someone who sets out to develop and sell a medicinal tea which relieves hangover symptoms. Overall success is selling your tea and making money. Investing in optimizing reality would mean investing in experiments to develop and improve the tea. Introducing variants, randomized controlled trials, etc., etc. Optimizing story means having a really good explanation for why your tea is able to do what you claim it does, plus having a great website with good copy, publishing testimonials, publishing your methodology and experimental results, etc.

Success can be attained by both, in combination, but also possibly each on its own. If you have a tea which works well, word will spread and you'll end up with many customers seeking your genuinely palliative tea. Alternatively, if your marketing materials are persuasive, you might accrue many customers too, even if your tea is no better than placebo. Of course, truly effective tea plus a well-conveyed story about its great properties will generate more sales than effective tea or a good story alone.

Optimizing Reality

The following points are worth noting:

- When optimizing reality, success depends on reality at large being a certain way.
- There is only one reality when it comes to what you're optimizing for.
- You can't fool reality. It is what it is no matter what you say or do.
- Optimizing for reality is easiest to do when outcomes are easy to measure and feedback is quick.
- Engineers dealing with concrete, measurable phenomena are likely to be optimizing reality.
- Genuinely creating value, though often expensive, is a good strategy for capturing value to yourself. Likely it's the best long-term strategy in many

domains.
    - I would claim that every major successful company is producing clear value to someone. Facebook, Google, AirBnB, ExxonMobil, United Airlines, Starbucks. At the end of the day, there's something real there.

Optimizing Stories

- For any goal, there might be many possible stories which will lead to success.
- Stories don't have to be true or accurate to be effective.
- It can be rational, or outright necessary, for one's success to optimize for stories without regard to reality.
    - If your success ultimately depends on someone else being convinced, then reality at large doesn't matter.
- Salespeople, marketers, politicians, startup founders and generally those whose success depends on persuading others will spend much of their effort optimizing stories.
- Very often your "customer" only cares that you have a "good story" and not at all that your story matches reality.
    - Consider a reseller of your medicinal tea. If they're unscrupulous, it doesn't matter to them whether or not the tea works. It only matters to them that your story is good enough to persuade end-consumers.
    - Consider a hospital IT employee whose job is to purchase software which helps doctors. The supposed goal of improved doctor efficiency is not measured in any way. The hospital IT employee's success will be judged by the impressive-seemingness of the software they select - how good the story it comes with - not any actual reality. In other words, the IT employee, as a customer, is incentivized to be shopping for good stories and only good stories.
        - But you needn't assume people only purchase stories for the sake of others! Many people, having a loose relationship, are happy to purchase a story which makes them feel good. "Oh, those crystals from the mystery mountains have the right frequencies to bring me good fortune? Yes please!"
- In cases where those hearing the stories lack the expertise to assess reality, e.g. non-experts trying assess an expert, "dumbed-down" stories comprehensible to the non-experts will completely outweigh reality itself. *Cf.* Overconfident talking down, humble or hostile talking up.
- If it's only stories which matter, yet you split your efforts between stories and reality, then you will likely be outcompeted by someone who spent all of their resources on crafting good stories. *Cf. Moloch.*
- Even those who care a lot about reality itself can slide into a focus on stories if empirical feedback is absent or very slow.
- Notwithstanding all of the above, stories will sometimes reach the end of their tether and the lack of a good reality to support your stories will catch up with you.
- We don't just tell stories to others, we also tell themselves to ourselves.
    - The danger of being too much in the habit of telling stories is that we don't merely risking fooling others, but also ourselves.

"Story Economies"

I conjecture that what arises in our modern world are “*economies” of stories* whereby people buy and sell stories often without regard for reality.

Another example: imagine an analyst working at a startup crafts a report which highlights all the ways in which the company is rapidly improving. The analyst's manager isn’t too worried about whether the report is a bit biased towards the positive - they know the CEO will be pleased. The CEO doesn’t mind if the report is a bit biased towards the positive - they know the board will be pleased. The board doesn’t mind if the report is a bit biased - they know that the next round of investors won’t really be able to tell the difference, it will just make the company look good.

Here you have a whole chain of people who only care about the story. At the very end there’s someone who cares about the reality, but they’re very often not in a great position to evaluate it themselves. They probably don’t know even the right questions to ask. All they’ve got is the story which has been placed before them.

Some people gravitate more towards stories than others, e.g. salespeople and politicians. Some of them might readily admit that they chiefly deal in stories somewhat tenuously linked to reality, yet I wager that many, if not most, won’t. The most persuasive stories are those you devoutly believe yourself. Hence the vast overconfidence of startup founders. And, in the immortal words of George Costanza: *it's not a lie... if you believe it.*

Stories about yourself ...to yourself and others

If there’s one domain where we’re endlessly crafting and broadcasting stories, it’s the stories we tell about ourselves. I’m *this* kind of person. I might decide that the story I want to tell is that I’m a "science nerd". So I read science books and science magazines. I have my answer ready when people ask me what I do for fun and I know exactly what to post on social media. My Instagram is full of homemade volcanoes and photos from my personal backyard telescope.

The above fictional example might have the redeeming feature that at least this fictional person is creating a genuine reality to match the story. They are learning a tonne of science genre facts. Still, I wager there’s a tradeoff. Doing science-y things which are easily communicable and demonstrable introduces a constraint. Possibly leveling up as a scientist would mean reading textbooks with facts that are incomprehensible and boring to those not at that level. By trying to have the best story to tell, they’ve handicapped their own excellence. (However, if the story is primarily for oneself, this constraint is avoided. “*I know just how science-y I am!)*

If you want to know, I tell myself the story that I’m a person who’s afraid of losing myself to trying craft myself into someone optimized for impressing others. Though I do it. I’m doing it right now. There may be no escape.

Cf. Elephant in the Brain.

You can’t escape stories

At this point you might be thinking, “gee, stories are awfully deceitful and non-cooperate, I want to cooperative and honest and I’m just going to provide direct facts!” and “I really, really don’t want to deceive myself with stories!”

I don’t think you can escape stories entirely. I would claim that as soon as you summarize your facts or data, the mere selection of which facts to present or summarize is the crafting of a story. Even dumping all your data and every observation is likely to be biased by which data you collected and what you paid attention to. What you thought were the relevant things to report to another person.

That said, I think there’s storytelling which attempts to be honest effort to share reality as is so that someone else can make an informed judgment. It’s challenging if one’s success is threatened by less scrupulous competitors, but it’s possible to choose domains where measurable feedback favors those who’ve optimized actual reality.

You’re not always doing others a favor if you try to give them raw facts with no biased conclusions. The world is large and messy and confusing such that people usually like to be handed a story about who you are and how you will behave. They want you to be a nerdy, bookish type, or an outdoorsy type, or a foodie. If you give me a story and promise to act in accordance with it, that makes simple. It’s clear what to talk about, what to get you for your birthday, etc., etc.

At least for those spending much time out in mainstream culture, it helps to have one or two stories prepared about yourself. “Masks.” They function a bit like APIs, really. People often protest that they don’t like being put in boxes, but those boxes help you relate to people before you’ve spent the many, many hours to have absorbed the messy reality that any given human is.

What to do, what to do

Reality on the ground is complex, incentives are messy, things which work in the short run don’t necessarily work in the long run. I can't say “here’s my one simple recipe to determine the right allocation of resources between optimizing story vs optimizing reality.”

I proffer the obvious advice:

- Notice the incentives for each domain you deal in - how much does your success depend on stories vs direct reality?
- By judging the incentives in the domains you deal, assess how much you can judge the stories you are presented with.
- Accept that the tradeoffs are hard. Personally, I wish I could deal only in truth and provide only open and transparent facts. Unfortunately, I might need either to compromise or to find myself severely disadvantaged in certain arenas, e.g. politics.
    - Moreover, the incentives mean despite myself, self-interest bias means I’m likely to present things to others in ways which favor myself.
- Accept that you probably need a story about yourself. If you like, keep the story separate from yourself so that you might let yourself be more. *Cf.* *Keep your identity small*.

How much can value learning be disentangled?

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

In the context of whether the definition of human values can disentangled from the process of approximating/implementing that definition, David asks me:

- But I think it's reasonable to assume (within the bounds of a discussion) that there is a non-terrible way (in principle) to specify things like "manipulation". So do you disagree?

I think it's a really good question, and its answer is related to a lot of relevant issues, so I put this here as a top-level post. My current feeling is, contrary to my previous intuitions, that things like "manipulation" might not be possible to specify in a way that leads to useful disentanglement.

Why manipulate?

First of all, we should ask why an AI would be tempted to manipulate us in the first place. It may be that it needs us to do something for it to accomplish its goal; in that case it is trying to manipulate our actions. Or maybe its goal includes something that cashes out as out mental states; in that case, it is trying to manipulate our mental state directly.

The problem is that any reasonable friendly AI would have our mental states as part of its goal - it would at least want us to be happy rather than miserable. And (almost) any AI that wasn't perfectly indifferent to our actions would be trying to manipulate us just to get its goals accomplished.

So manipulation is to be expected by most AI designs, friendly or not.

Manipulation versus explanation

Well, since the urge to manipulate is expected to be present, could we just rule it out? The problem is that we need to define the difference between manipulation and explanation.

Suppose I am fully aligned/corrigible/nice or whatever other properties you might desire, and I want to inform you of something important and relevant. In doing so, especially if I am more intelligent than you, I will simplify, I will omit irrelevant details, I will omit arguably relevant details, I will emphasise things that help you get a better understanding of my position, and de-emphasise things that will just confuse you.

And these are exactly the same sorts of behaviours that smart manipulator would do. Nor can we define the difference as whether the AI is truthful or not. We want human understanding of the problem, not truth. It's perfectly possible to manipulate people while telling them nothing but the truth. And if the AI structures the order in which it

presents the true facts, it can manipulate people while presenting the whole truth as well as nothing but the truth.

It seems that the only difference between manipulation and explanation is whether we end up with a better understanding of the situation at the end. And measuring understanding is very subtle. And even if we do it right, note that we have now motivated the AI to... aim for a particular set of mental states. We are rewarding it for manipulating us. This is contrary to the standard understanding of manipulation, which focuses on the means, not the end result.

Bad behaviour and good values

Does this mean that the situation is completely hopeless? No. There are certain manipulative practices that we might choose to ban. Especially if the AI is limited in capability at some level, this would force it to follow behaviours that are less likely to be manipulative.

Essentially, there is no boundary between manipulation and explanation, but there is a difference between extreme manipulation and explanation, so ruling out the first can help (or maybe not).

The other thing that can be done is to ensure that the AI has values close to ours. The closer the values of the AI are to us, the less manipulation it will need to use, and the less egregious the manipulation will be. It might be that, between partial value convergence and ruling out specific practices (and maybe some physical constraints), we may be able to get an AI that is very unlikely to manipulate us much.

Incidentally, I feel the same about low-impact approaches. The full generality problem, an AI that is low impact but value-agnostic, I think is impossible. But if the values of the AI are better aligned with us, and more physically constrained, then low impact becomes easier to define.

[Speech] Worlds That Never Were

This is the speech I gave at the 2018 Bay Area Secular Solstice. It was written to fit the bill of 'concrete visions of the future.' There is one small change here from the version I gave : the Solstice organizers had me change the second-to-last sentence of the speech, because the event centered around the idea of disasters that fall short of extinction, and rebuilding from the ashes. I've changed it back to be more true to what I think Solstice is about - the complete and irrecoverable disaster looming on the horizon, and the reality of our desperate efforts to avoid it.

There are lots of people on the Internet and on TV every day telling you how the world is going to dissolve into chaos at any moment. When you think about all the people and animals needlessly suffering, it's easy to think that the world is terrible. When you read the news, or look at the current state of the art in preparedness for global catastrophic risks, it's easy to think that the world is doomed, and feel hopeless. But that's why we're here. We are here – in this community and in this planetarium – because we believe that there is a future worth fighting for.

What does that future look like? That's hard to say. You know how sometimes, when you look directly at a star in the night sky, it fades so that you can barely see it, but when you look a little to the side it's bright again? That's sort of how I feel about imagining the future. It's slippery, hard to get a grasp on, sort of just out of reach. But for now, let's try to picture it.

Imagine the world a few hundred years from now. Humanity has completed a process of great deliberation. We've overcome global coordination problems, so there's no more poverty and no more war, and existential risks have shrunk to a millionth of a millionth of their current size. We have overcome physics and engineering challenges and have ships on their way to the distant reaches of the galaxy. We have overcome our biological limitations and no longer die so easily of disease or accident or old age.

If we sit with that, it still feels abstract, and far away. So try to imagine yourself in that future, even if how we imagine it today will never be how it really is.

Maybe you're with your family on a spaceship, off to colonize the edge of the galaxy. Your faces are pressed to the window as you watch the Earth grow small against the blackness. Though you're leaving this planet behind, probes have been sent out to terraform the barren rocks that lie ahead, and they'll be lush with life when you arrive. When the pale blue dot has faded to nothing in the distance, you turn away from the window and get to work preparing for the long journey ahead. Your children call up to you and you answer with confidence. You're in this together. You believe you will make it to your destination. You are not afraid.

Or maybe you're there on the day the last factory farm closes. Advances in lab-grown meat and our understanding of nutrition have obviated the need for such cruelty, and our new ethical standards would never allow it. You watch the animals stumble, confused, into the daylight. It's too late for most of them, but at least they are the last. Humanity will never again commit torture like this.

Or maybe you work for the government, which is now a legible institution whose mission feels like your mission. Every day you go to work and uphold a system of laws

that you believe in, and every night you return home to your children, who are educated in a system that actually lets them learn, while keeping their spirits bright and alive. Around the dinner table at night, they debate with you about the ethics of terraforming, tell you the history of the great deliberation, and explain to you concepts in physics that you could never quite grasp. Your bodies are warm, your stomachs are full, and your minds are active. You're free to pursue whatever you wish, without the constant nagging guilt that you could be doing more to help others and prevent extinction, because that's all been taken care of.

The path ahead is full of obstacles. We don't yet have the technology to build space colonies or cheat old age, nor have we figured out what incentives will end war, poverty, or factory farming for good. No matter how hard we work, there's always a chance we won't make it to tomorrow. But humanity has proved time and again that we can overcome forces of nature once thought to be unstoppable, so no matter how slim the odds, there's also a chance that we'll survive. And if we do, we should hold onto those images of how good the future could be.

By and large, the people here tonight are not religious. We know that we live in a world beyond the reach of God. If we are going to make it out of this, it will have to be by our own power. If the future is going to be as good as the stories I've told, we have to make it that way. We only have one chance. Let's make tomorrow beautiful.

What is a reasonable outside view for the fate of social movements?

Epistemic status: very hand-wavy and vague, but confident there is a substantial and well-understood core. Hoping for an answer that elucidates that core more clearly.

It is something of a rationalist folk theorem that social movements face the risk of an "Eternal September", or of scaling into oblivion. (See e.g. this blog post by Leverage research, this paper by Owen Cotton-Barratt, David Chapman or Benjamin Hoffman on "Geeks, MOPs and sociopaths", and Scott Alexander on "the toxoplasma of rage").

I've had the sense that some cocktail of Hansonian/Dunbarian evolutionary psychology, basic game theory/microeconomics, memetic theory and Sturgeon's law, should predict this. That is, that some reasonably operationalised version of the claim "most social movements fail" is true.

Yet I am not able to point to >=5 historical examples of social movements that suffered this fate, along with some gears for what went wrong.

Hence I'm looking for links, historical examples, more fleshed-out gears, ... anything that might form a more rigorous reference point for *an outside view on the fate of social movements*.

Which approach is most promising for aligned AGI?

I know that this is something of a speculative question and that people will have wildly different views here, but it seems important to try to have an idea of which approaches are more likely than others to lead to AGI. It's okay to argue that multiple approaches are promising, but you might want to consider seperate answers if you have a reasonable amount to write about each approach.

Is Agent Simulates Predictor a "fair" problem?

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

It's a simple question, but I think it might help if I add in context. In the paper introducing Functional Decision Theory, it is noted that it is impossible to design an algorithm that can perform well on all decision problems since some of them can be specified to be blatantly unfair, ie. punish every agent that isn't an alphabetical decision theorist.

The question then arises, how do we define which problems are or are not fair? We start by noting that some people consider Newcomb's-like problems to be unfair since your outcome depends on a predictor's prediction, which is rooted in an analysis of your algorithm. So what makes this case any different from only rewarding the alphabetical decision theorist?

The paper answers that the prediction only depends on the decision you end up making and that any other internal details are ignored. So it only cares about your decision and not how you come to it, the problem seems fair. I'm inclined to agree with this reasoning, but a similar line of reasoning doesn't seem to hold with Agent Simulates Predictor. Here the algorithm you use is relevant as the predictor can only predict the agent if it's algorithm is less than a certain level of complexity, otherwise it may make a mistake.

Please note that this question isn't about whether this problem is worth considering; life is often unfair and we have to deal with it the best that we can. The question is about whether the problem is "fair", where I roughly understand "fair" meaning that this is in a certain class of problems that I can't specify at this moment (I suspect it would require its own seperate post) where we should be able to achieve the optimal result in each problem.

What math do i need for data analysis?

I always had fairly good mathematical thinking (I think) and loved learning about beautiful concepts in math - but i didn't learn much at all in school (cause i had the choice). You can say i was "utilitarian" regarding learning math, i didn't do it if i didn't see how it can enrich my life.

so my knowledge of math is quite disorganized, i know more about Bayes theorem then many much simpler concepts (i know, it really shouldn't be that way).

Now i want to be able to analyze data, but i don't want to learn math that i won't use for it, if possible.

So here's my question - what basic stuff do i need to learn in order to be able to calculate probabilities, statistics, do Bayesian math, and overall do things within data analysis that I may yet be aware of?

If you also have suggestions for how to learn *those* things, after i learn the basics, it will be much appreciated.

thank you :)

What are the components of intellectual honesty?

I have a pretty strong intuition that "intellectual honesty" points to a specific thing, with a characteristic set of behaviors (both outward (in a conversation/writing) and inward (patterns of thinking)). My concept of intellectual honesty also seems to largely coincide with the concept as used by others, but I'm not sure.

I am talking about a high standard. This isn't at all about violating a code of academic ethics.

I am talking about the sort of thing which helps to foster epistemic trust, helping to take conversations to higher levels. (But don't read those links if you already know enough what I mean to answer and want to avoid anchoring your view to mine.)

One Website To Rule Them All?

Epistemic status: I'm not the first person to think of this, and I'm not sure if this is possible. But I want to flesh out some options for consideration and get feedback. Sorry if it's long.

WANTED:

A crowd-sourced site that reliably presents important pros and cons on any/all topics and hopefully leads us collectively closer to consensus.

This has been on my mind for a few years now. Figuring out truth in the digital age is often tricky and sometimes downright impossible. (Is minimum wage a good thing? Should I adopt a paleo or keto or vegan or Shangri-la diet? What do we really know and not know about the historical Jesus?) Especially when it comes to political discussion, answers often depend on more details than an average person has time or energy to research—and even if they did, they may still leave out considerations or solutions that someone else may think of.

Some links where other rationalists think semi-related things: Robin Hanson wants AI to help us towards consensus because Wikipedia isn't up for the job. Scott Alexander promoted adversarial collaborations to see if they could reach consensus. (I swear, I saw more examples of people wishing for something like a debate website, but I wasn't writing them down and searching for them has failed.)

THE WISHFUL THINKING PART

Now let me be more specific about what my imagined website (hereafter, Site) needs to do.

On matters of t*ruth,* it needs to support epistemic arguments for w*hy w*e should believe or not believe particular claims. On matters of *action,* it needs to provide important pro/cons of taking that action. Site must have a method of allowing the best arguments to rise to the top. That's the most basic functionality.

Additional Feature A: Because the Site is crowd-sourced, anyone can propose possible actions and possible modifications to actions. In theory, this could mean that Site could be a platform where a crowd-written law could have its details hammered out by many affected parties. Creative solutions could have a chance to be reviewed and, if well-thought-out, could rise higher in the public consciousness.

Optional Feature B: I really, really would like it if, within the structure of arguments (for either epistemic-truth topics or proposed-actions), Site distinguished between truth-claims and value-claims. This might be especially unrealistic, since no one naturally lays out arguments that way. But gosh dang it, I really want it to.

Optional Feature C: A number of rationalist posts have discussed how useful it would be for society to have a way to coordinate on certain problems: specifically, having a conditional "I will adopt X if Y% of other relevant people also do". It might be possible to set up Site so that proposed actions that need a minimum number of people to adopt it in order for it to be worthwhile offer the option of conditionally committing to adopt it and sending out notifications when thresholds are reached.

THE NITTY-GRITTY

Obviously, this is a LOT to ask from a website. Before I go into everything that could go wrong and all the reasons it might not even be possible, let me lay out a couple of operational options.

Contrast Voting

Reddit and other up/down-vote systems suffer from certain effects, an important one being that late-to-the-game comments don't stand a chance against early popular comments, even if the later ones are higher quality. To combat this, Site should not allow up/down votes on individual items/comments/whatever. Instead, Site should present *two* items and ask for a vote on *which one is better.* In this way, Site can give better rankings to items. A new argument, perhaps one that presents a new, more recent study finding, can quickly outrank the previous top argument if, when those two arguments are presented, most people vote that the new one is better.

Contrast Topics

The Contrast Topics option would not have a page titled "Is God real?" nor a page for "We should outlaw chicken cage farming." Instead, topics would be presented in terms of possible alternatives. "The Abrahamic God is real" could be separately contrasted with "There is no God" and, on another page, contrasted with "There is a divine energy". "We should outlaw chicken cage farming" could be contrasted with "We should eat less chicken and let the market adjust chicken cage farming downward" or "We should slowly phase out chicken cage farming according to this detailed plan I've written out" or "We should wait until chicken-like lab meat is available, and then outlaw chicken cage farming".

Visually, this might be tricky, but a "Chicken cage farming" page might list all of the proposed actions related to it, and a user could select two actions and click "next" to see the pros and cons of those two actions relative to each other. (These would be more like pros of A vs pros of B, not pro/cons of each; a con of B would be listed as a pro of A and vice versa.) These headings might be treated as tags/labels rather than a strict hierarchy of categories, so that a possible action or topic might appear under multiple headings if it's relevant to all of them. Headings might, in turn, have both epistemic truth-claims and potential actions listed under them.

Open Forum

In this option, individual users compose complete arguments, either pro or con. They are allowed and even encouraged to steal what others have written, add/subtract/modify it as they like, and submit it as a new argument. Contrast Voting is used within "pro" and within "con" to rank these arguments and present the highest ranked at the top. Probably only the top one or two submissions will show by default (the rest reached by an expandable + or "show more" or such), since submissions are expected to (eventually) aggregate all the best arguments.

Wikibate

This is an alternative to Open Forum. It works more like Wikipedia; everyone is allowed to directly edit one single page. A pro and con side are provided, and we see if a version comes out that makes enough people happy enough to be stable. A behind-the-scenes contrast voting option might be used to decide which order the possible

arguments appear in, but the phrasing of the arguments would have to satisfy people or be edited by someone.

Personal Expertise Stories

In normal life, people rely heavily on personal experiences when forming their beliefs. If Site doesn't sometimes take this into account, it will feel shallow to many people. Argue the facts of abortion all you want, but if you can't include someone's vivid description of having a loved one die from a back-alley abortion or another person's horror at seeing what they were told is a "lump of tissue" that turns out to look exactly like a baby, you're missing important context.

It is also the case that sometimes hearing expert testimony is helpful. Lawmakers often rely on it. Experts don't always agree, but a claim from an expert can and should carry a different weight than the same claim from a non-expert.

On a crowd-sourced site, anyone can claim they're an expert or make up b.s. stories about what happened to them. A possible solution is to allow people to submit a "personal expertise/story". If submitting one, you have to include personal contact info like a phone number. Some number, say 10, people can apply to verify a story. Only approved verifiers are given access to the story submitter's contact info. The verifiers promise to come back and report on the status of their verification and not to abuse the contact info. (A complaint from a story submitter results in that verifier being unable to do any further verifications, and possibly banned from the Site.) The verifier calls the person and talk to them, get details if possible, like other people or companies they can talk to, to verify parts of the story or claim. In the end they report back, maybe on a 5-point scale, whether they believe the person's claim, and the net verification result from all the verifiers is displayed with the story. (E.g. a green bar showing 3.5/5, with a (2) next to it indicating only 2 verifiers have submitted responses.)

Consequences Before Arguments

For proposed actions, *before* the actual pro/con arguments appear, I would like to see a list of possible consequences that might result from taking that action. While I don't want Site to adopt consequentialism as an official philosophy, I remember (Google, you fail me) an article that claimed that people demonstrated less partisanship when they were asked to think about the consequences of possible laws. A potential consequence is itself a truth-claim ("If we do X, then Y will happen"), so that should link to its own page.

Login Voting

Generally, I want people not to have to create accounts in order to participate on the website. IP addresses can be recorded, like Wikipedia does, for submitting/modifying arguments, proposing actions, and creating topics. That should minimize the barrier to entry and encourage participation. However, all voting should require logging in, to minimize one person voting multiple times, and to accurately track if someone changes their relative rankings of two arguments.

Parallel Axes Voting

There is no one standard for what makes an argument "better" than another. Relevant measurements include: accuracy, importance/applicability, thoroughness, kindness, formatting/grammar/spelling. One option might be to allow users to Contrast Vote two

arguments in multiple categories separately: Argument A has more important points than B, but B has better grammar than A, and they are the same on kindness. Parallel Axes Voting is more relevant if Site uses the Open Forum style, and not so much if it uses Wikibate.

THE PROBLEMS, OH SO MANY PROBLEMS

- **I don't know how to make this website.** I know a little html and css, and I have a vague idea that Amazon Web Services could be initially used to host the site. But I know nothing about how to use AWS, how to make a database and a website talk to each other, how to have user accounts and secure logins, etc. I'd either need a lot of help or someone else would have to do it entirely or I'd need to invest a ton of time into basically learning a new profession to make something that might fail.
- **The most popular arguments aren't necessarily the best.**
    - *People might sacrifice truth for simplicity.* Given two arguments that make the same point, one that presents the point more simply and understandably is better than one that doesn't. However, users will sometimes be offered the choice between a simple, understandable argument that isn't accurate and a difficult one that is more correct. I don't know of a way to discourage upvoting the simpler one over the more accurate one. Worse, if all the arguments on the inaccurate side are simple and easy to understand, and all the arguments on the more correct side are difficult to understand, Site could backfire and have the effect of convincing people of something that isn't true.
    - *Some topics require a whole background course to understand.* Israel/Palestine, the mortgage crisis...there's some things that require so much background knowledge (even to be familiar with the relevant terms) that I'm not sure it's possible to create arguments that the average reader could understand without also presenting some sort of background course on the topic. [Maybe this could be ameliorated by limiting the voting on certain topics to people who have accessed and agreed that they have read a background page that verified experts have approved?]
- **I don't know if Site will actually promote consensus.** I think seeing your opponent's best arguments instead of only their worst will reduce partisanship a little. I think that comparing multiple possible actions on a topic instead of only one or two will help some. But that might not be enough.
    - Having all arguments divided into either pro or con instead of one single narrative might not shift people into a consensus. Relatedly, it might be difficult for the two sides to adequately interact with each other and respond to each other's views when everything is listed on one side or the other.
    - Seeing the pros and cons of both sides might actually make untrue beliefs (flat Earth, for example) seem more legitimate or reasonable than they are; the epistemic imbalances might not come across effectively.
    - Much disagreement includes trusting sources differently. If an argument about whether something did/would happen or not boils down to "Breitbart/Slate said so", Site might not have a way to resolve that.
    - A website might be innately insufficient; personally caring about another person you know might be required for consensus.
    - A sub-point on this one is that in the Wikibate style, I don't know if arguments can reach a reasonably stable state. Wikipedia does sometimes have those background pages for arguing about whether an article should say one thing or another; Wikibate might try those for settling disputes,

deciding whether subtle differences are similar enough to count as the same thing or not, and such. But contentious things are contentious, and that might not be enough.

- **Some algorithms and UI details need to be worked out.**
    - *How can the conditional-cooperation option be implemented?* If different people have different thresholds that need to be met before they will do it, how does Site take that into account? How does it handle identifying the subpopulation of the whole planet that would need to conditionally commit to certain measures? (E.g. all corporations will follow some environmental rule if all the others do too—the CEOs or other higher-ups in the corporations would be the only relevant population that would need to meet the threshold, not every individual on the planet).
    - *What formula should Site use for taking user's Contrast Votes and turning them into an overall rank?* What happens when people be their inconsistent selves and like A better than B, B better than C, but C better than A?
    - *How do we get users to rank new submissions?* Do we list new ones separately above the already-ranked ones? Force a pop-up that asks people to rank a newer submission with a random older one before they continue reading that page? Offer a sidebar that says "Here's some new arguments, please review and rank them:"?
    - *Numbers.* This is especially relevant when listing possible consequences of actions, but sometimes Site needs to handle numbers delicately. "One more person each year will die if we do this" is different from "100,000 more people will die each year if we do this". Listing all the possible estimates that individuals pull out of their...err, hat would be overwhelming to look at and consider. Do we create a separate page for arguments over what the number will be and institute Contrast Voting on the results so that the original page shows the most popular estimate? List upper and lower bounds: "Between 1 and 100,000 people will die"?
    - *How does Site handle nesting?* How far down do we let nesting continue before deciding that two options are too similar to bother listing them separately? Does Site make you compare only the lowest-nested level with each other, or you compare higher nested levels too? (That is, is there a page for Abrahamic God vs. No God and also a page for prayer-answering Baptist God vs. No God, even though prayer-answering Baptist God is nested under Abrahamic God? Or can you only compare prayer-answering Baptist God to non-so-interventionist Baptist God? If so, how do you get agreement on what aspects of God all the Abrahamics agree on?)
    - *Is a summary possible?* Would it be possible for a page to present a sort of consensus summary? (Something like "People prefer immediate outlawing of chicken cage farming if they value reducing the total amount of suffering of living creatures more than they value reducing human suffering alone, and people prefer gradual phasing out of chicken cage farming if they value reducing human suffering alone more than they value reducing the total amount of suffering of living creatures".)
    - *How does Site balance covering all the options with presenting a number of possibilities that people will actually read?* Sometimes it will be enough to list all the options, put the most important/relevant ones at the top, and let people read as far down as they want. Sometimes that might not be enough. For instance, if you want to list the consequences of a possible action before you list the pro/con arguments for that action, then what do you do if there's 50 consequences proposed? Cut it off at an arbitrary number (say, only show the 5 most important)? Display any that, say, 50%

of people vote should be visible without being hidden behind a "see more" option?
    - *How is moderation handled?* Like, I have no clue. This probably varies depending on whether Site uses Open Forum or Wikibate.
    - There's a lot more details to work out. I think I have some implicit imagery in my head as to how Site looks and operates that I might not have laid out explicitly here, but a lot of it has yet to be worked out. And doubtless huge new problems will be encountered once actually trying to make the site.
- **Getting Site to be popular enough to be useful is difficult.** I definitely don't have the social network or following to pull something like that off. I don't really know anyone who does. The best I could do is email a link (once it's set up) to random famous people and say "Please check this out". Maybe if someone writes up a "6 Reasons to Check Out This Website (And Laugh)" article and submits it to Buzzfeed.
- **Funding.** If it does become popular, decisions will need to be made about advertising and how to pay for server maintenance and whatnot.
- **Site needs a cool name.** All I've come up with is Wikibate (if it's set up that way) or BetterThink.

In the end, despite the problems, I think Site would be more worth having than not.

So what do you think? Is Site even possible? Is there a better setup than Open Forum or Wikibate? If not, which of those two is better? Should the Personal Expertise Stories option be included? What other details could Site implement to be successful? Are any of you willing to work on it? Is it worth my time to work on it?

Comments on CAIS

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

Over the last few months I've talked with Eric Drexler a number of times about his Comprehensive AI Services (CAIS) model of AI development, and read most of his technical report on the topic. I think these are important ideas which are well worth engaging with, despite personally being skeptical about many of the conclusions. Below I've summarised what I see as the core components of Eric's view, followed by some of own arguments. Note that these are only my personal opinions. I did make some changes to the summary based on Eric's comments on early drafts, to better reflect his position - however, there are likely still ways I've misrepresented him. Also note that this was written before reading Rohin's summary of the same report, although I do broadly agree with most of Rohin's points.

One useful piece of context for this model is Eric's background in nanotechnology, and his advocacy for the development of nanotech as "atomically precise manufacturing" rather than self-replicating nanomachines. The relationship between these two frameworks has clear parallels with the relationship between CAIS and a recursively self-improving superintelligence.

The CAIS model:

1. The standard arguments in AI safety are concerned with the development of a single AGI agent doing open-ended optimisation. Before we build such an entity (if we do so at all), we will build AI services which each perform a bounded task with bounded resources, and which can be combined to achieve superhuman performance on a wide range of tasks.
2. AI services may or may not be "agents". However, under CAIS there will be no entity optimising extremely hard towards its goals in the way that most AI safety researchers have been worrying about, because:
    1. Each service will be relatively specialised and myopic (focused on current episodic performance, not maximisation over the whole future). This is true of basically all current AI applications, e.g. image classifiers or Google Translate.
    2. Although rational agents can be proved equivalent to utility-maximisers, the same is not necessarily true of systems of rational agents. Most such systems are fundamentally different in structure from rational agents - for example, individual agents within the system can compete with or criticise each other. And since AI services aren't "rational agents" in the first place, a system composed of them is even less likely to implement a utility-maximiser.
    3. There won't be very much demand for unified AIs which autonomously carry out large-scale tasks requiring general capabilities, because systems of AI services will be able to perform those tasks just as well or better.
3. Early AI services could do things like massively disrupt financial markets, increase the rate of scientific discovery, help run companies, etc. Eventually they should be able to do any task that humans can, at our level or higher.
    1. They could also be used to recursively improve AI technologies and to develop AI applications, but usually with humans in the loop - in roughly

the same way that science allows us to build better tools with which to do better science.

4. Our priorities in doing AI safety research can and should be informed by this model:
    1. A main role for technical AI safety researchers should be to look at the emergent properties of systems of AI services, e.g. which combinations of architectures, tasks and selection pressures could lead to risky behaviour, as well as the standard problems of specifying bounded tasks.
    2. AI safety experts can also give ongoing advice and steer the development of AI services. AI safety researchers shouldn't think of safety as a one-shot problem, but rather a series of ongoing adjustments.
    3. AI services will make it much easier to prevent the development of unbounded agent-like AGI through methods like increasing coordination and enabling surveillance, if the political will can be mustered.

I'm broadly sympathetic to the empirical claim that we'll develop AI services which can replace humans at most cognitively difficult jobs significantly before we develop any single superhuman AGI (one unified system that can do nearly all cognitive tasks as well as or better than any human). One plausible mechanism is that deep learning continues to succeed on tasks where there's lots of training data, but doesn't learn how to reason in general ways - e.g. it could learn from court documents how to imitate lawyers well enough to replace them in most cases, without being able to understand law in the way humans do. Self-driving cars are another pertinent example. If that pattern repeats across most human professions, we might see massive societal shifts well before AI becomes dangerous in the adversarial way that's usually discussed in the context of AI safety.

If I had to sum up my objections to Eric's framework in one sentence, it would be: "the more powerful each service is, the harder it is to ensure it's individually safe; the less powerful each service is, the harder it is to combine them in a way that's competitive with unified agents." I've laid out my arguments in more detail below.

Richard's view:

1. Open-ended agentlike AI seems like the most likely candidate for the first strongly superhuman AGI system.
    1. As a basic prior, our only example of general intelligence so far is ourselves - a species composed of agentlike individuals who pursue open-ended goals. So it makes sense to expect AGIs to be similar - especially if you believe that our progress in artificial intelligence is largely driven by semi-random search with lots of compute (like evolution was) rather than principled intelligent design.
        1. In particular, the way we trained on the world - both as a species and as individuals - was by interacting with it in a fairly unconstrained way. Many machine learning researchers believe that we'll get superhuman AGI via a similar approach, by training RL agents in simulated worlds. Even if we then used such agents as "services", they wouldn't be bounded in the way predicted by CAIS.
    2. Many complex tasks don't easily decompose into separable subtasks. For instance, while writing this post I had to keep my holistic impression of Eric's ideas in mind most of the time. This impression was formed through having conversations and reading essays, but was updated frequently as I wrote this post, and also draws on a wide range of my background knowledge. I don't see how CAIS would split the task of understanding a

high-level idea between multiple services, or (if it were done by a single service) how that service would interact with an essay-writing service, or an AI-safety-research service.
    1. Note that this isn't an argument against AGI being modular, but rather an argument that requiring the roles of each module and the ways they interface with each other to be human-specified or even just human-comprehensible will be very uncompetitive compared with learning them in an unconstrained way. Even on today's relatively simple tasks, we already see end-to-end training outcompeting other approaches, and learned representations outperforming human-made representations. The basic reason is that we aren't smart enough to understand how the best cognitive structures or representations work. Yet it's key to CAIS that each service performs a specific known task, rather than just doing useful computation in general - otherwise we could consider each lobe of the human brain to be a "service", and the combination of them to be unsafe in all the standard ways.
    2. It's not clear to me whether this is also an argument against IDA. I think that it probably is, but to a lesser extent, because IDA allows multiple layers of task decomposition which are incomprehensible to humans before bottoming out in subtasks which we can perform.
3. Even if task decomposition can be solved, humans reuse most of the same cognitive faculties for most of the tasks that we can carry out. If many AI services end up requiring similar faculties to each other, it would likely be more efficient to unify them into a single entity. It would also be more efficient if that entity could pick up new tasks in the same rapid way that humans do, because then you wouldn't need to keep retraining. At that point, it seems like you no longer have an AI service but rather the same sort of AGI that we're usually worried about. (In other words, meta-learning is very important but doesn't fit naturally into CAIS).
4. Humans think in terms of individuals with goals, and so even if there's an equally good approach to AGI which doesn't conceive of it as a single goal-directed agent, researchers will be biased against it.

2. Even assuming that the first superintelligent AGI is in fact a system of services as described by the CAIS framework, it will be much more like an agent optimising for an open-ended goal than Eric claims.
    1. There'll be significant pressure to reduce the extent to which humans are in the loop of AI services, for efficiency reasons. E.g. when a CEO can't improve on the strategic advice given to it by an AI, or the implementation by another AI, there's no reason to have that CEO any more. Then we'll see consolidation of narrow AIs into one overall system which makes decisions and takes actions, and may well be given an unbounded goal like "maximise shareholder value". (Eric agrees that this is dangerous, and considers it more relevant than other threat models).
    2. Even if we have lots of individually bounded-yet-efficacious modules, the task of combining them to perform well in new tasks seems like a difficult one which will require a broad understanding of the world. An overseer service which is trained to combine those modules to perform arbitrary tasks may be dangerous because if it is goal-oriented, it can use those modules to fulfil its goals (on the assumption that for most complex tasks, some combination of modules performs well - if not, then we'll be using a different approach anyway).
        1. While I accept that many services can be trained in a way which makes them naturally bounded and myopic, this is much less clear to

me in the case of an overseer which is responsible for large-scale allocation of other services. In addition to superhuman planning capabilities and world-knowledge, it would probably require arbitrarily long episodes so that it can implement and monitor complex plans. My guess is that Eric would argue that this overseer would itself be composed of bounded services, in which case the real disagreement is how competitive that decomposition would be (which relates to point 1.2 above).

3. Even assuming that the first superintelligent AGI is in fact a system of services as described the CAIS framework, focusing on superintelligent agents which pursue unbounded goals is still more useful for technical researchers. (Note that I'm less confident in this claim than the others).
    1. Eventually we'll have the technology to build unified agents doing unbounded maximisation. Once built, such agents will eventually overtake CAIS superintelligences because they'll have more efficient internal structure and will be optimising harder for self-improvement. We shouldn't rely on global coordination to prevent people from building unbounded optimisers, because it's hard and humans are generally bad at it.
    2. Conditional on both sorts of superintelligences existing, I think (and I would guess that Eric agrees) that CAIS superintelligences are significantly less likely to cause existential catastrophe. And in general, it's easier to reduce the absolute likelihood of an event the more likely it is (even a 10% reduction of a 50% risk is more impactful than a 90% reduction of a 5% risk). So unless we think that technical research to reduce the probability of CAIS catastrophes is significantly more tractable than other technical AI safety research, it shouldn't be our main focus.

As a more general note, I think that one of the main strengths of CAIS is in forcing us to be more specific about what tasks we envisage AGI being used for, rather than picturing it divorced from development and deployment scenarios. However, I worry that the fuzziness of the usual concept of AGI has now been replaced by a fuzzy notion of "service" which makes sense in our current context, but may not in the context of much more powerful AI technology. So while CAIS may be a good model of early steps towards AGI, I think it is a worse model of the period I'm most worried about. I find CAIS most valuable in its role as a research agenda (as opposed to a predictive framework): it seems worth further investigating the properties of AIs composed of modular and bounded subsystems, and the ways in which they might be safer (or more dangerous) than alternatives.

Many thanks to Eric for the time he spent explaining his ideas and commenting on drafts. I also particularly appreciated feedback from Owain Evans, Rohin Shah and Jan Leike.

Why don't people use formal methods?

This is a linkpost for https://www.hillelwayne.com/post/why-dont-people-use-formal-methods/

Saw this on Hacker News; there's Hacker News discussion here.

Formal methods seem very relevant to AI safety, and I haven't seen much discussion of them on Less Wrong.

Littlewood's Law and the Global Media

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

"Forecasting Transformative AI: An Expert Survey", Gruetzemacher et al 2019

This is a linkpost for https://arxiv.org/abs/1901.08579

Prediction Contest 2018: Scores and Retrospective

Way back in April 2018, I announced a Prediction Contest, in which the person who made the best predictions on a bunch of questions on PredictionBook ahead of a 1st July deadline would win a prize after they all resolved in January 2019, which is now.

It was a bit of an experiment; I had no idea how many people were up for practicing predictions to try to improve their calibration, and decided to throw a little money and time at giving it a try. And in the spirit of reporting negative experimental results: The answer was 3, all of which I greatly appreciate for their participation. I don't regret running the experiment, but I'm going to pass on running a Prediction Contest 2019. I don't think this necessarily rules out trying to practically test and compete in rationality-related areas in other ways later, though.

The Results

Our entrants were bendini, bw, and Ialaithion, and their ranked log scores were:

bw: -9.358221122

Ialaithion: -9.594999615

bendini: -10.0044594

This was sufficiently close that changing a single question's resolution could tip the results, so they were all pretty good. That said, bw came out ahead, and even managed to beat averaging everyone's predictions- if you simply took the average prediction (including non-entrants) as of entry deadline and made that your prediction, you'd have got -9.576568147.

The full calculations for each of the log scores, as well as my own log score and the results of feeding the predictions as of prediction time to a simple model rather than simply averaging them, are in a spreadsheet here.

I'll be in touch with bw to sort out their prize this evening, and thanks to everyone who participated and who helped with finding questions to use for it.