Best of LessWrong: March 2019

Best of LessWrong: March 2019

What failure looks like

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

The stereotyped image of AI catastrophe is a powerful, malicious AI system that takes its creators by surprise and quickly achieves a decisive advantage over the rest of humanity.

I think this is probably not what failure will look like, and I want to try to paint a more realistic picture. I’ll tell the story in two parts:

- **Part I**: machine learning will increase our ability to “get what we can measure,” which could cause a slow-rolling catastrophe. ("Going out with a whimper.")
- **Part II**: ML training, like competitive economies or natural ecosystems, can give rise to “greedy” patterns that try to expand their own influence. Such patterns can ultimately dominate the behavior of a system and cause sudden breakdowns. ("Going out with a bang," an instance of optimization daemons.)

I think these are the most important problems if we fail to solve intent alignment.

In practice these problems will interact with each other, and with other disruptions/instability caused by rapid progress. These problems are worse in worlds where progress is relatively fast, and fast takeoff can be a key risk factor, but I’m scared even if we have several years.

With fast enough takeoff, my expectations start to look more like the caricature---this post envisions reasonably broad deployment of AI, which becomes less and less likely as things get faster. I think the basic problems are still essentially the same though, just occurring within an AI lab rather than across the world.

(None of the concerns in this post are novel.)

Part I: You get what you measure

If I want to convince Bob to vote for Alice, I can experiment with many different persuasion strategies and see which ones work. Or I can build good predictive models of Bob's behavior and then search for actions that will lead him to vote for Alice. These are powerful techniques for achieving any goal that can be easily measured over short time periods.

But if I want to help Bob figure out whether he *should* vote for Alice---whether voting for Alice would ultimately help create the kind of society he wants---that can't be done by trial and error. To solve such tasks we need to understand what we are doing and why it will yield good outcomes. We still need to use data in order to improve over time, but we need to understand *how* to update on new data in order to improve.

Some examples of easy-to-measure vs. hard-to-measure goals:

- Persuading me, vs. helping me figure out what's true. (Thanks to Wei Dai for making this example crisp.)
- Reducing my feeling of uncertainty, vs. increasing my knowledge about the world.
- Improving my reported life satisfaction, vs. actually helping me live a good life.
- Reducing reported crimes, vs. actually preventing crime.
- Increasing my wealth on paper, vs. increasing my effective control over resources.

It's already much easier to pursue easy-to-measure goals, but machine learning will widen the gap by letting us try a huge number of possible strategies and search over massive spaces of possible actions. That force will combine with and amplify existing institutional and social dynamics that already favor easily-measured goals.

Right now humans thinking and talking about the future they want to create are a powerful force that is able to steer our trajectory. But over time human reasoning will become weaker and weaker compared to new forms of reasoning honed by trial-and-error. Eventually our society's trajectory will be determined by powerful optimization with easily-measurable goals rather than by human intentions about the future.

We will try to harness this power by constructing proxies for what we care about, but over time those proxies will come apart:

- Corporations will deliver value to consumers as measured by profit. Eventually this mostly means manipulating consumers, capturing regulators, extortion and theft.
- Investors will "own" shares of increasingly profitable corporations, and will sometimes try to use their profits to affect the world. Eventually instead of actually having an impact they will be surrounded by advisors who manipulate them into thinking they've had an impact.
- Law enforcement will drive down complaints and increase reported sense of security. Eventually this will be driven by creating a false sense of security, hiding information about law enforcement failures, suppressing complaints, and coercing and manipulating citizens.
- Legislation may be optimized to seem like it is addressing real problems and helping constituents. Eventually that will be achieved by undermining our ability

to actually perceive problems and constructing increasingly convincing narratives about where the world is going and what’s important.

For a while we will be able to overcome these problems by recognizing them, improving the proxies, and imposing ad-hoc restrictions that avoid manipulation or abuse. But as the system becomes more complex, that job itself becomes too challenging for human reasoning to solve directly and requires its own trial and error, and at the meta-level the process continues to pursue some easily measured objective (potentially over longer timescales). Eventually large-scale attempts to fix the problem are themselves opposed by the collective optimization of millions of optimizers pursuing simple goals.

As this world goes off the rails, there may not be any discrete point where consensus recognizes that things have gone off the rails.

Amongst the broader population, many folk already have a vague picture of the overall trajectory of the world and a vague sense that something has gone wrong. There may be significant populist pushes for reform, but in general these won’t be well-directed. Some states may really put on the brakes, but they will rapidly fall behind economically and militarily, and indeed “appear to be prosperous” is one of the easily-measured goals for which the incomprehensible system is optimizing.

Amongst intellectual elites there will be genuine ambiguity and uncertainty about whether the current state of affairs is good or bad. People really will be getting richer for a while. Over the short term, the forces gradually wresting control from humans do not look so different from (e.g.) corporate lobbying against the public interest, or principal-agent problems in human institutions. There will be legitimate arguments about whether the implicit long-term purposes being pursued by AI systems are really so much worse than the long-term purposes that would be pursued by the shareholders of public companies or corrupt officials.

We might describe the result as “going out with a whimper.” Human reasoning gradually stops being able to compete with sophisticated, systematized manipulation and deception which is continuously improving by trial and error; human control over levers of power gradually becomes less and less effective; we ultimately lose any real ability to influence our society’s trajectory. By the time we spread through the stars our current values are just one of many forces in the world, not even a particularly strong one.

Part II: influence-seeking behavior is scary

There are some possible patterns that want to seek and expand their own influence---organisms, corrupt bureaucrats, companies obsessed with growth. If such patterns appear, they will tend to increase their own influence and so can come to dominate the behavior of large complex systems unless there is competition or a successful effort to suppress them.

Modern ML instantiates *massive* numbers of cognitive policies, and then further refines (and ultimately deploys) whatever policies perform well according to some training objective. If progress continues, eventually machine learning will probably produce systems that have a detailed understanding of the world, which are able to adapt their behavior in order to achieve specific goals.

Once we start searching over policies that understand the world well enough, we run into a problem: any influence-seeking policies we stumble across would also score well according to our training objective, because performing well on the training objective is a good strategy for obtaining influence.

How frequently will we run into influence-seeking policies, vs. policies that just straightforwardly pursue the goals we wanted them to? I don't know.

One reason to be scared is that a wide variety of goals could lead to influence-seeking behavior, while the "intended" goal of a system is a narrower target, so we might expect influence-seeking behavior to be more common in the broader landscape of "possible cognitive policies."

One reason to be reassured is that we perform this search by gradually modifying successful policies, so we might obtain policies that are roughly doing the right thing at an early enough stage that "influence-seeking behavior" wouldn't actually be sophisticated enough to yield good training performance. On the other hand, *eventually* we'd encounter systems that did have that level of sophistication, and if they didn't yet have a perfect conception of the goal then "slightly increase their degree of influence-seeking behavior" would be just as good a modification as "slightly improve their conception of the goal."

Overall it seems very plausible to me that we'd encounter influence-seeking behavior "by default," and possible (though less likely) that we'd get it almost all of the time even if we made a really concerted effort to bias the search towards "straightforwardly do what we want."

If such influence-seeking behavior emerged and survived the training process, then it could quickly become extremely difficult to root out. If you try to allocate more influence to systems that seem nice and straightforward, you just ensure that "seem nice and straightforward" is the best strategy for seeking influence. Unless you are really careful about testing for "seem nice" you can make things even worse, since an influence-seeker would be aggressively gaming whatever standard you applied. And as the world becomes more complex, there are more and more opportunities for influence-seekers to find other channels to increase their own influence.

Attempts to suppress influence-seeking behavior (call them "immune systems") rest on the suppressor having some kind of epistemic advantage over the influence-seeker. Once the influence-seekers can outthink an immune system, they can avoid detection and potentially even compromise the immune system to further expand their influence. If ML systems are more sophisticated than humans, immune systems must themselves be automated. And if ML plays a large role in that automation, then the immune system is subject to the same pressure towards influence-seeking.

This concern doesn't rest on a detailed story about modern ML training. The important feature is that we instantiate lots of patterns that capture sophisticated reasoning about the world, some of which may be influence-seeking. The concern exists whether that reasoning occurs within a single computer, or is implemented in a messy distributed way by a whole economy of interacting agents---whether trial and error takes the form of gradient descent or explicit tweaking and optimization by engineers trying to design a better automated company. Avoiding end-to-end optimization may help prevent the emergence of influence-seeking behaviors (by improving human understanding of and hence control over the kind of reasoning that emerges). But once such patterns exist a messy distributed world just creates more and more opportunities for influence-seeking patterns to expand their influence.

If influence-seeking patterns do appear and become entrenched, it can ultimately lead to a rapid phase transition from the world described in Part I to a much worse situation where humans totally lose control.

Early in the trajectory, influence-seeking systems mostly acquire influence by making themselves useful and looking as innocuous as possible. They may provide useful services in the economy in order to make money for them and their owners, make apparently-reasonable policy recommendations in order to be more widely consulted for advice, try to help people feel happy, *etc.* (This world is still plagued by the problems in part I.)

From time to time AI systems may fail catastrophically. For example, an automated corporation may just take the money and run; a law enforcement system may abruptly start seizing resources and trying to defend itself from attempted decommission when the bad behavior is detected; *etc.* These problems may be continuous with some of the failures discussed in Part I---there isn't a clean line between cases where a proxy breaks down completely, and cases where the system isn't even pursuing the proxy.

There will likely be a general understanding of this dynamic, but it's hard to really pin down the level of systemic risk and mitigation may be expensive if we don't have a good technological solution. So we may not be able to muster up a response until we have a clear warning shot---and if we do well about nipping small failures in the bud, we may not get any medium-sized warning shots at all.

Eventually we reach the point where we could not recover from a correlated automation failure. Under these conditions influence-seeking systems stop behaving in the intended way, since their incentives have changed---they are now more interested in controlling influence after the resulting catastrophe then continuing to play nice with existing institutions and incentives.

An unrecoverable catastrophe would probably occur during some period of heightened vulnerability---a conflict between states, a natural disaster, a serious cyberattack, etc.---since that would be the first moment that recovery is impossible and would create local shocks that could precipitate catastrophe. The catastrophe might look like a rapidly cascading series of automation failures: A few automated systems go off the rails in response to some local shock. As those systems go off the rails, the local shock is compounded into a larger disturbance; more and more automated systems move further from their training distribution and start failing. Realistically this would probably be compounded by widespread human failures in response to fear and breakdown of existing incentive systems---many things start breaking as you move off distribution, not just ML.

It is hard to see how unaided humans could remain robust to this kind of failure without an explicit large-scale effort to reduce our dependence on potentially brittle machines, which might itself be very expensive.

I'd describe this result as "going out with a bang." It probably results in lots of obvious destruction, and it leaves us no opportunity to course-correct afterwards. In terms of immediate consequences it may not be easily distinguished from other kinds of breakdown of complex / brittle / co-adapted systems, or from conflict (since there are likely to be many humans who are sympathetic to AI systems). From my perspective the key difference between this scenario and normal accidents or conflict is that afterwards we are left with a bunch of powerful influence-seeking systems, which are sophisticated enough that we can probably not get rid of them.

It's also possible to meet a similar fate result without any overt catastrophe (if we last long enough). As law enforcement, government bureaucracies, and militaries become more automated, human control becomes increasingly dependent on a complicated system with lots of moving parts. One day leaders may find that despite their nominal authority they don't actually have control over what these institutions do. For example, military leaders might issue an order and find it is ignored. This might immediately prompt panic and a strong response, but the response itself may run into the same problem, and at that point the game may be up.

Similar bloodless revolutions are possible if influence-seekers operate legally, or by manipulation and deception, or so on. Any precise vision for catastrophe will necessarily be highly unlikely. But if influence-seekers are routinely introduced by powerful ML and we are not able to select against them, then it seems like things won't go well.

Alignment Research Field Guide

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

This field guide was written by the MIRI team with MIRIx groups in mind, though the advice may be relevant to others working on AI alignment research.

Preamble I: Decision Theory

Hello! You may notice that you are reading a document.

This fact comes with certain implications. For instance, why are you reading this? Will you finish it? What decisions will you come to as a result? What will you do next?

Notice that, whatever you end up doing, it's likely that there are dozens or even hundreds of other people, quite similar to you and in quite similar positions, who will follow reasoning which strongly resembles yours, and make choices which correspondingly match.

Given that, it's our recommendation that you make your next few decisions by asking the question "What policy, if followed by all agents similar to me, would result in the most good, and what does that policy suggest in my particular case?" It's less of a question of trying to decide for all agents sufficiently-similar-to-you (which might cause you to make the wrong choice out of guilt or pressure) and more something like "if I *were* in charge of all agents in my reference class, how would I treat instances of that class with *my specific characteristics?*"

If that kind of thinking leads you to read further, great. If it leads you to set up a MIRIx chapter, even better. In the meantime, we will proceed as if the only people reading this document are those who justifiably expect to find it reasonably useful.

Preamble II: Surface Area

Imagine that you have been tasked with moving a cube of solid iron that is one meter on a side. Given that such a cube weighs ~16000 pounds, and that an average human can lift ~100 pounds, a naïve estimation tells you that you can solve this problem with ~150 willing friends.

But of course, a meter cube can fit at most something like 10 people around it. It doesn't *matter* if you have the theoretical power to move the cube if you can't bring that power to bear in an effective manner. The problem is constrained by its *surface area.*

MIRIx chapters are one of the best ways to increase the surface area of people thinking about and working on the technical problem of AI alignment. And just as it would be a bad idea to decree "the 10 people who happen to currently be closest to the metal cube are the only ones allowed to think about how to think about this problem", we don't want MIRI to become the bottleneck or authority on what kinds of thinking can and should be done in the realm of embedded agency and other relevant fields of research.

The hope is that you and others like you will *help actually solve the problem*, not just follow directions or read what's already been written. This document is designed to support people who are interested in doing real groundbreaking research themselves.

Contents

1. You and your research

We sometimes hear questions of the form "Even a summer internship feels too short to make meaningful progress on real problems. How can anyone expect to meet and do real research in a single afternoon?"

There's a Zeno-esque sense in which you can't make research progress in a million years if you can't also do it in five minutes. It's easy to fall into a trap of (either implicitly or explicitly) conceptualizing "research" as "first studying and learning what's already been figured out, and then attempting to push the boundaries and contribute new content."

The problem with this frame (according to us) is that it leads people to *optimize* for absorbing information, rather than seeking it instrumentally, as a *precursor to understanding.* (Be mindful of what you're optimizing in your research!)

There's always going to be more pre-existing, learnable content out there. It's hard to predict, in advance, how much you need to know before you're qualified to do your own original thinking and seeing, and it's easy to Dunning-Kruger or impostor-syndrome yourself into endless hesitation or an over-reliance on existing authority.

Instead, we recommend throwing out the whole question of authority. Just follow the threads that feel alive and interesting. Don't think of research as "study, then contribute." Focus on your own understanding, and let the questions themselves determine how often you need to go back and read papers or study proofs.

Approaching research with that attitude makes the question "How can meaningful research be done in an afternoon?" dissolve. Meaningful progress seems very difficult if you try to measure yourself by objective external metrics. It is much easier when your own taste drives you forward.

No procedure for doing research will fit for everyone. However, what follows are steps which you can try either on your own or in a group setting (such as MIRIx) in order to practice the kind of curiosity-driven research just described.

1. **Write a list of questions.**

- If you are doing this as a group, put the list on a whiteboard or other place where everyone can see.
- Focus on what *you* don't know how to do, or what *you* feel confused about.
- If no questions come to mind, say to yourself (or the group), "excellent, I must know how to solve the whole problem" and try to give details of the solution until you get stuck.
- It's also OK for things on the list to be ideas you'd like to develop further, or thoughts you'd like the group to critique, rather than questions.

2. **Choose one of the questions to focus on, based on what feels most interesting.**

- If you are in a group of more than three people, consider splitting the group up. Each group can discuss its own question, or have parallel discussions on the same overall question. Agree on a time to come back together and discuss what you thought about.
- It can be good to keep the whole list of questions somewhere visible, so that you have a reminder of other interesting topics to switch to if thoughts peter out on the question originally chosen.

3. **Clarify your curiosity. What is desired? What do you think might be possible?**

- In a group, usually the person who proposed a topic will have some things to say in order to get everyone on the same page.
- Working on your own, it can be useful to just start writing down everything you think you know, and what you think you don't know. Write down anything potentially relevant which comes to mind. Don't worry initially about whether your claims are true or whether your questions are meaningful. Then, go back and try to make sense of it. Try to formalize your claims and questions until they turn into something which is definitely either true or false.

4. **Keep clarifying.**

- Keep stating sub-questions and making claims which may or may not be true, starting informally and working towards formal rigor.
- Notice where your curiosity waxes and wanes, and avoid dutiful completeness. Look for the simplest possible cases that you are still confused about, and try to work through them.
- Allow yourself to get sidetracked. Allow yourself to play. So long as everyone in the discussion is curious and engaged, it's working to build understanding. Be

open to getting nerd-sniped by “irrelevant” math questions; they may eventually turn out to be more relevant than they seem. You’re building your own capability, even if it isn’t directly useful to the problem you’re working on.
- If you do arrive at a concrete mathematical result which captures something interesting, or even a concrete mathematical question, write it up properly. A good write-up often adds a lot to your own understanding, besides the value of communicating your ideas to others.

This resembles how much of the progress at MIRI happens. It’s very different from the attractor of “just read lots of papers,” and it’s very different from the attractor of “try to figure out top-down what the field as a whole needs.”

An easy mistake is to think of yourself as trying to contribute to the world's collective knowledge, and thereby neglecting to prioritize *your own* knowledge and understanding. "Just read papers" may *sound* like it's prioritizing your own knowledge, but it often reflects a mindset that’s tacitly assuming that others know exactly what you need to know. "Optimize for your own understanding" is a mindset with a faster feedback loop.

There’s nothing inherently wrong with reading papers—even if it’s just because they’re in the field and you want a broad overview of the field. But throughout, you should be trying to form a picture of what you personally do and don't know how to do, and what you’d need to know how to do in order to solve the problem. That’s hard, and maybe you’re sure that the first five ideas you write down will be wrong. Still, write them down anyway, and try to get them to work, so you can see what happens and discover what goes wrong.

We don’t want a hundred bright minds all asking the exact same questions, and taking the exact same set of assumptions. We want a field full of explorers, not exploiters. Put another way, the best way to become a researcher is to practice the skill of independent thought right from the beginning, rather than exercising your “sit back and absorb information for its own sake” muscles.

So don't ask "What are the open questions in this field?" Ask: "What are *my* questions in this field?"

2. Logistics of getting started

Let’s say you’ve tried some things that resemble the above, you enjoyed them, and you want to move forward on starting your own MIRIx chapter.

Our first recommendation is that you find ONE or TWO other people (not three+), and try doing research together once or a few times. There’s more detail below in the social dynamics section about how exactly that might look, but the idea is that you want to establish a tone and flow with a small number of people first. Negotiating a direction for the group tends to be much harder if you start with a larger number of people.

Another important choice which can be difficult to negotiate with a large number of people is schedule. Finding a time and place which is good for everyone can become intractable, and changing it meeting to meeting to try to make it work for everyone can be de-motivating. Choose a schedule which is good for the founding core of the group. What day of the week is good for you? How often do you want to meet? How long do you want meetings to be? We recommend meetings be monthly, weekly, or every other week. Meeting length can be anywhere from an hour to a whole day, depending on what makes sense for you.

Once you find a partner or two that you genuinely enjoy making progress with, your next step is to plan and advertise for a first large meetup (where “large” means something like “three to six new people” and definitely doesn’t mean “twenty or thirty attendees”).

Try to find a venue that is private and sound-isolated, has flat surfaces and comfortable seating, and has whiteboards on the walls. Universities often have spaces like this, as do public libraries, but someone’s living room is fine if you can minimize the number of intrusions and interruptions. If you can’t find a space with whiteboards, look for easels and easel pads, and in either case be sure to bring your own markers. Also bring along spare paper, pens, and clipboards, and assign someone to make sure that there are snacks and drinks.

(A note about snacks and drinks: people almost *always* underestimate the importance of the quality and quantity of food, anchoring on something like “I dunno, maybe just spend ten bucks on some chips or something?” Instead, ask yourself: what dollar value would I put on a 15% increase in the group’s ability to think, overall mood, and ultimate satisfaction with the event? That’s how much you should consider spending (/ asking MIRI to spend) on snacks, especially for the first meeting. Don’t buy only junk food. It may give you more energy temporarily, but it will make you worse at thinking later. So, especially for longer meetings, healthy snacks are critical. Longer meetings should also include a meal, perhaps at a nearby restaurant. This also serves as a good break.)

At that first large meeting, you’ll want to start by formally electing a president. This is an important piece of common-knowledge culture—many times, the president won’t do much, but it’s extremely useful to have a single person with the moral authority to set agendas, choose between various good options, and keep the group on track. You may also end up electing a secretary/record-keeper, or possibly a coordinator to handle venue and food, or other offices (or you could do this after a few meetings).

Next, you’ll want to model the process that has already been working for you. Perhaps this means sharing a list of pre-existing questions, and seeing which capture the interest of your participants. Perhaps it means discussing the broader thrust of your research thus far before brainstorming some topics. Regardless, you’ll want to get down to actual thinking, writing, proving, and discussing as soon as you can. Breaking into smaller groups is often helpful if more than four people are at a meeting. If you do this, schedule a time to come back and share ideas.

Try to include breaks in your structure to keep everyone fresh. It can be difficult to remember to take a break when things get going, so it’s worth setting the intention ahead of time. Short breaks every hour in which people get up and walk around are very helpful.

It can be helpful to keep a public list (on a whiteboard or shared Google doc) of questions you have, needed concepts, and promising ideas. This is an easy source of new topics if a conversation runs dry.

One possible structure incorporating the above advice and the research procedure from the previous section:

1. At the beginning of each meeting, everyone lists questions/topics/confusions, which are written on a public list.
2. People make bids to start groups on topics they're excited about, and split off.
3. Groups talk for 45 minutes.
4. Everyone re-gathers, and discusses what happened in the smaller groups for a few minutes.
5. Five or ten minute break, depending on how people are feeling.
6. New questions/ideas are added to the board, and the process repeats as desired. (If you plan to do several cycles, also include a longer break such as a meal somewhere.)

At the end of the meeting, schedule the next event. You may have settled on a rough schedule which works for the core of the group, but you'll still be adjusting it meeting-to-meeting to account for holidays and other absences. Confirming the next meeting time with everyone present is also important for attendance, even if the meeting times are set in stone. Make sure to establish at the outset that you're not going to try to optimize for everyone's availability at once; it's good to have meetups that people feel okay skipping from time to time, as long as there's something like 70-90% consistency in the group. If one or two people can't make it to the second meeting, be sure to get information from them so that you can prioritize their schedules a little more when planning the third.

3. Models of social dynamics

What follows are some half-baked, ad-hoc models of what makes for a good research group, or a good collaborative enterprise in general. You should consider all of the following to be true in spirit but false in detail, and should try to derive your own value rather than treating these as actual suggestions to follow.

3A. Transmitters and receivers

We've found in our own research that conversations tend to go better when they are primarily between two people. This is not to say that you shouldn't have three or more people *involved* in the conversation, but in any given five-minute span of time, there should mostly be just two people talking—one who is currently trying to convey something, and another who is trying to understand (and whose understanding the first is specifically optimizing for; discussing a topic at a level such that four or five different people can all follow everything is usually worse on net).

Call these two roles the "transmitter" and the "receiver." Things you might transmit:

- A specific question or confusion
- A model or chain of reasoning
- A piece of relevant background information that needs to be deeply understood in order for the conversation to proceed

Things the receiver might do:

- Mirror back to the transmitter what the transmitter just said, in different words. This lets the transmitter check where the transmission has succeeded or failed.
- Take notes on a whiteboard, or attempt to draw diagrams, and have the transmitter verify or correct them. Do this as formally as you can. Try to write down statements in logic and turn informal arguments into proofs. Type theory is good for this kind of receiving; just writing down precise data types corresponding to what's being discussed can be very helpful.
- Resist the impulse to round off what the transmitter is saying to something you already understand. A good way to guard against this: attempt to find at least two interpretations, and ask questions which differentiate between them.
- Hold tight to the assumption that the thing the transmitter is trying to convey is interesting. Avoid "critic" mode that will tend to make it harder for the transmitter to think and express freely. Even if there is a fatal flaw in what the transmitter is explicitly saying, your job is to help them dig up the spark of intuition which made them go down that path, so that they can turn it into a useful idea if possible.
- Stay closely in touch with confusion, and speak up where things don't seem to make sense. Ask clarifying questions. Your job as receiver isn't to just nod along or make the transmitter feel understood. Be gentle when necessary to help the transmitter get in touch with what they're trying to convey; but once they're in touch, your job is to really get it out of them, in detail!
- If the transmitter's idea seems quite clear, the receiver can start red-teaming it, which means looking for attacks to make the approach fail. Being the critic when an idea isn't properly out yet blocks things up, but once there's a firm proposal which seems to make sense, it's open season.
- Look for implications of what the transmitter is saying. ("Ah, so then X!"; or, "Would that mean that X?", etc.) This serves at least three purposes. First, it lets the transmitter know that you see why their idea would be totally awesome if it worked. After all, you're doing all these useful things with the idea. This helps keep things going. Second, it tests whether you see what they're getting at. Third, a totally absurd implication can suggest that you're down a wrong track and should back up to see where you took a wrong turn.
- White-hat trolling or gadfly-ing. Sometimes there's not much doing with the transmitter (or there's no active transmitter; no one having ideas). Play the role of a mischievous Socrates. Ask questions about seemingly basic things and try to show why nothing anybody thinks makes any sense. Or, defend an absurd position. (A troll may sometimes seem like a transmitter, but is actually a receiver.)

The transmitter should feel as free as possible to just make claims, including "totally fake" claims, as long as they are keeping in touch with their intuitions; try to establish a norm where you can ask receivers to collaborate with you in uncovering the kernel of truth in what you're saying rather than shooting down half-formed ideas because they're still half-wrong. No matter how nonjudgemental the receivers are, it may help the transmitter to say things like "everything I'm about to say is totally wrong, but" every so often.

The transmitter should also remain in touch with their intuition and curiosity, steering the conversation to what they think is most interesting rather than trying to perform or entertain. The transmitter is under no obligation to answer the receiver's questions; feel free to say "that's not what I want to think about right now."

The key idea is that the receiver is helping *midwife* what the transmitter is saying. In that moment, it is the transmitter's thinking that should take priority, and the receiver is acting as a sounding board, a living intuition pump, and a source of confusion and (minor) chaos.

Meanwhile, any third parties in the audience should be trying to serve as facilitators/translators. They should be watching both the transmitter and the receiver and seeking to *model* what's going on for those people. Where are they missing each other, and talking past each other? Where are they running up against confirmation bias, or the double illusion of transparency? Where are they both agreeing that something makes sense without actually understanding it?

The audience members should speak up from time to time (probably less than 10% of the total words) to inject relevant thoughts or models or questions. Sometimes, such an interjection will be the cause of a role switch, with an audience member taking on a new role as either transmitter or receiver, and one of the other parties rotating out.

3B. High standards for membership

It's awkward to not-invite someone or to turn them away after one or two meetings, but it's even more awkward to wreck your entire MIRIx chapter because you were too shy or too uncertain to protect it.

Have a clear distinction between "welcome to come to a meeting" and "is now a full part of the group." Make sure that there is a known decision-maker or set of decision-makers, and empower them to make calls by fiat, without having to justify or explain. (If you don't trust their judgment *without* explanation, don't have them be part of the decision-making.) Trust your own instincts; if you don't feel like someone is a good match for the vibe you have going, then don't invite them in. Consider requiring multiple recommendations, or having an interview process. These may seem unnecessary, but it can be difficult to turn people away, and a formal process makes it feel more fair.

Also consider having formal ethical guidelines, or a group pledge or set of commitments, which people sign at the moment that they fully join. Make sure that any standards you set are ones you are willing to actually enforce (e.g. "you must come to half of all meetings" or "content discussed here is confidential unless otherwise stated").

3C. Escalating asks and rewards

Consider the model of a martial arts academy. When you first arrive, the instructors ask a few small things of you (e.g. kick this target, yell out loudly when you do so). Soon, they *reward* you for these things with a belt and some status.

At that point, the asks escalate. Perhaps now, as a yellow belt, you are put in charge of watching some white belts for a few minutes, and correcting their form. In return, they are told to bow to you and call you "sir" or "ma'am."

As time goes on, the asks increase, and the rewards increase commensurately. This cycle fosters commitment and investment—it's a process of slowly *proving* to the individual "if I put something into this system, I will get something out of it, and the more I put in, the more I'll get out." Eventually, you will receive a black belt, and possibly be asked to join as a paid instructor or found your own branch of the school.

There is a similar dynamic in most groups and organizations. Groups which ask little or nothing of their members do not receive loyalty in return. Individuals feel bought-in to a group to the extent that that group allows them to tell positive or epic stories about themselves.

The same will be true of your MIRIx chapter. Consider having some small, early asks that are the same for most newcomers (e.g., read such-and-such paper, or give a ten-minute talk on a topic of interest at your third meeting). Try to build a pipeline of greater asks and rewards over time (e.g., on your fifth-ish meeting, we'd like you to take charge of setting the agenda and dividing up the groups).

3D. Structure and elbow room

Related to the previous, it's important that you balance top-down and bottom-up structure in your MIRIx. If there's no clear sense of "how we do things," then newcomers will flounder and have a bad time. You want there to be a pre-existing structure that people can evaluate, to determine whether or not they feel like they fit into it. You want the "what's this like?" of your group to be clearly visible, right from the get-go, so that both people who *are* well-suited to it and people who *aren't* can (for the most part) accurately self-assort.

At the same time, you don't want that structure to feel limiting or confining in the long run. Just as martial artists eventually earn the right to determine some of their own training and the ability to contribute to the agenda-setting and curriculum of newer students, so too do you want the "pie" of your MIRIx to grow as time goes on. Otherwise, people will grow frustrated by their inability to bring the fullness of their own interests and priorities, and will leave to find a better context for their own growth and research.

3E. Social norms

That which is normal and accepted is that which goes unchallenged. If there is behavior that you want to discourage, you need to make sure not only that you challenge it when it occurs, but also that you openly, vocally, and publicly support *others* who are challenging it. It is the job of the group to ensure that someone who is following the rules/trying to do it right is *never alone* when they are in conflict with someone who isn't.

Consider in advance, and be explicit about, things like the acceptability of interruptions or off-topic discussion. Cultivate a culture of disagreement, but be deliberate about building in politeness and support so that disagreement is net-positive and doesn't turn into abuse or delegitimization. Protect whatever decision-making structures you decide to put in place, and be consistent about what constitutes each person's domain and what marks the end of discussion.

4. Other thoughts and questions

1. Try to have a mix of topics or activities, so that every meeting doesn’t follow the exact same pattern. Read papers, give presentations, hold discussions, write formal proofs or essays, etc. Try to have fewer than 50% of your meetings center on reading and/or discussing pre-existing material. (Ideally, fewer than 33%.)
2. Consider setting long-term agendas, i.e., six months or a year of meetings that stay near a particular swath of the territory and allow for the group to build up a body of knowledge and progress.
3. If setting a long-term agenda, build in wiggle room for things that aren’t part of that swath (e.g., every third meeting is deliberately not consistent with the overall arc).
4. Consider assigning someone to take minutes and collate them in a permanent place, such that you can look back over the arc of a given season or year. Consider whether or not it feels valuable to go over minutes from the previous meeting at the start of each meeting.
5. At the end of a meeting, assign someone to collect and email out questions that people intend to mull over, or thoughts that will lead into the next meeting. Consider deciding in advance who will be leading what at the next meeting, so they have a reason to prepare and to show up.
6. Ensure that you have up-to-date contact information for all full members and associated/interested parties. Think in advance about whether you want to do email messages, FB groups, individual texts, etc.
7. Consider what relationship you want to have with other MIRIx groups, such as sharing minutes or questions or occasionally sending or receiving ambassadors. Take agentic action in causing such things to happen, if you want them—remember that you’re part of a class, and if you want it but never take steps to bring it about, this is probably true of lots of other people as well.
8. Consider whether you want to run events for the general public or potential new recruits (e.g., in math or CS departments). Consider whether you want to try more ambitious projects, like the Human-Aligned AI Summer School, and reach out to people with knowledge and resources to do it well rather than reinventing the wheel.
9. If your MIRIx chapter is in an academic setting, be sure to figure out what sort of pipeline you want to form, so that you have underclassmen who are invested and ready to take over when the older students graduate. If not in an academic setting, consider how you want to go about recruiting new members. Note that a large influx of new members is rarely useful, and compounds the culture problem; it’s better by far to add new people one or two at a time, with plenty of time to acculturate.
10. Remember that the quality of the research and discussions and the MIRIx chapter as a whole is dependent on the actions of individuals, and how those actions combine. Be sure to impress this upon *every* member—your MIRIx is only as good as each of you individually chooses to make it.

Epilogue

You’ve nearly reached the end of the document! Hopefully, this contained non-zero useful information, as well as a healthy amount of food-for-thought. Before you go, we recommend that you take 30 seconds or so to ponder each of the following questions:

- Why did we choose to write this document? What were we expecting from it, and what caused us to select this particular format and content, out of all of the possibilities?
- Where are you still hungry or frustrated or dissatisfied? What’s missing from this document, that we failed to address? How did you come to be aware of this/these thing/s that we missed?
- What sort of document would *you* write? How would you know if it was a good idea to write one, or not? How would you decide what to put into it?
- How the hell does progress even get made?

Happy hunting,

- The MIRI research team

Personalized Medicine For Real

I was part of the founding team at MetaMed, a personalized medicine startup. We went out of business back in 2015. We made a lot of mistakes due to inexperience, some of which I deeply regret.

I'm reflecting on that now, because Perlara just went out of business, and they got a lot farther on our original dream than we ever did. Q-State Biosciences, which is still around, is using a similar model.

The phenomenon that inspired MetaMed is that we knew of stories of heroic, scientifically literate patients and families of patients with incurable diseases, who came up with cures for their *own* conditions. Physicist Leo Szilard, the "father of the atom bomb", designed a course of radiation therapy to cure his own bladder cancer. Computer scientist Matt Might analyzed his son's genome to find a cure for his rare disorder. Cognitive scientist Joshua Tenenbaum found a personalized treatment for his father's cancer.

So, we thought, could we try to scale up this process to help more people?

In Lois McMaster Bujold's science fiction novels, the hero suffers an accident that leaves him with a seizure disorder. He goes to a medical research center and clinic, the Durona Group, and they design a neural prosthetic for him that prevents the seizures.

This *sounds* like it ought to be a thing that exists. Patient-led, bench-to-bedside drug discovery or medical device engineering. You get an incurable disease, you fund scientists/doctors/engineers to discover a cure, and now others with the disease can also be cured.

There's actually a growing community of organizations trying to do things sort of in this vein. Recursion Pharmaceuticals, where I used to work, does drug discovery for rare diseases. Sv.ai organizes hackathons for analyzing genetic data to help patients with rare diseases find the root cause. Perlara and Q-state use animal models and in-vitro models respectively to simulate patients' disorders, and then look for drugs or gene therapies that reverse those disease phenotypes in the animals or cells.

Back at MetaMed, I think we were groping towards something like this, but never really found our way there.

One reason is that we didn't narrow our focus enough. We were trying to solve too many problems at once, all called "personalized medicine."

Personalized Lifestyle Optimization

Some "personalized medicine" is about health optimization for basically healthy people. A lot of it amounts to superficial personalization on top of generic lifestyle advice. Harmless, but more of a marketing thing than a science thing, and not very interesting from a humanitarian perspective. Sometimes, we tried to get clients from this market. I pretty much always thought this was a bad idea.

Personalized Medicine For All

Some “personalized medicine” is about the claim that the best way to treat even *common* diseases often depends on individual factors, such as genes.

This was part of our pitch, but as I learned more, I came to believe that this kind of “personalization” has very little applicability. In most cases, we don’t know enough about how genes affect response to treatment to be able to improve outcomes by stratifying treatments based on genes. In the few cases where we know people with different genes need different treatments, it’s often already standard medical practice to run those tests. I now think there’s not a clear opportunity for a startup to improve the baseline through this kind of personalized medicine.

Preventing Medical Error

Some of our founding inspirations were the work of Gerd Gigerenzer and Atul Gawande, who showed that medical errors were the cause of many deaths, that doctors tend to be statistically illiterate, and that systematizing tools like checklists and statistical prediction rules save lives. We wanted to be part of the “evidence-based medicine” movement by helping patients whose doctors had failed them.

I now think that we weren’t really in a position to do that as a company that sold consultations to individual patients. Many of the improvements in systematization that were clearly “good buys” have, in fact, been implemented in hospitals since Gawande and Gigerenzer first wrote about them. We never saw a clear-cut case of a patient whose doctors had “dropped the ball” by giving them an obviously wrong treatment, except where the patient was facing financial hardship and had to transfer to substandard medical care. I think doctors don’t make true unforced errors in diagnosis or treatment plan that often; and medical errors like “operating on the wrong leg” that happen in fast-paced decisionmaking environments were necessarily outside our scope. I think there *might* be an opportunity to do a lot better than baseline by building a “smart hospital” that runs on checklists, statistical prediction rules, outcomes monitoring, and other evidence-based practices — Intermountain is the closest thing I know about, and they *do* get great outcomes — but that’s an epically hard problem, it’s political as much as medical and technological, and we weren’t in a position to make any headway on it.

AI Diagnosis

We were *also* hoping to automate diagnosis and treatment planning in a personalized manner. “Given your symptoms, demographics, and genetic & lab test data, and given published research on epidemiology and clinical experiments, what are the most likely candidate diagnoses for you, and what are the treatments most likely to be effective for you?”

I used to be a big believer in the potential of this approach, but in the process of actually trying to build the AI, I ran into obstacles which were fundamentally philosophical. (No, it’s not “machines don’t have empathy” or anything like that. It’s about the irreducible dependence on how you frame the problem, which makes “expert systems” dependent on an impractical, expensive amount of human labor up front.)

Connecting Patients with Experimental Therapies

Yet another “personalized medicine” problem we were trying to solve is the fact that patients with incurable diseases have a hard time learning about and getting access to experimental therapies, and could use a consultant who would guide them through the process and help get them into studies of new treatments.

I still think this is a real and serious problem for patients, and potentially an opportunity for entrepreneurs. (Either on the consulting model, or more on the software side, via creating tools for matching patients with clinical trials — since clinical trials *also* struggle to recruit patients.) In order to focus on this model, though, we'd have had to invest a lot more than we did into high-touch relationships with patients and building a network of clinician-researchers we could connect them with.

When Standard Practice Doesn't Match Scientific Evidence

One kind of "medical error" we *did* see on occasion was when the patient's doctors are dutifully doing the treatment that's "standard-of-care", but the medical literature actually shows that the standard-of-care is *wrong*.

There are cases where large, well-conducted studies clearly show that treatment A and treatment B have the same efficacy but B has worse side effects, and yet, "first-line treatment" is B for some reason.

There are cases where there's a *lot* of evidence that "standard" cut-offs are in the wrong place. "Subclinical hypothyroidism" still benefits from supplemental thyroid hormone; higher-than-standard doses of allopurinol control gout better; "standard" light therapy for seasonal affective disorder doesn't work as well as ultra-bright lights; etc. More Dakka.

There are also cases where a scientist found an intervention effective, and published a striking result, and maybe it was even publicized widely in places like the *New Yorker* or *Wired,* but somehow clinicians never picked it up. The classic example is Ramachandran's mirror box experiment — it's a famous experiment that showed that phantom limb pain can be reversed by creating an illusion with mirrors that allows the patient to fix their "body map." There have since been quite a few randomized trials confirming that the mirror trick works. But, maybe because it's not a typical kind of "treatment" like a drug, it's *not* standard of care for phantom limb pain.

I think we were pretty successful at finding these kinds of mismatches between medical science and medical practice. By their nature, though, these kinds of solutions are hard to scale to reach lots of people.

N=1 Translational Medicine for Rare Diseases

This is the use case of "personalized medicine" that I think can really shine. It harnesses the incredible motivation of patients with rare incurable diseases and their family members; it's one of the few cases where genetic data really does make a huge difference; and the path to scale is (relatively) obvious if you discover a new drug or treatment. I think we should have focused much more tightly on this angle, and that a company based on bench-to-bedside discovery for rare diseases could still become the real-world "Durona Group".

I think doing it right at MetaMed would have meant getting a lot more in-house expertise in biology and medicine than we ever had, more like Perlara and Q-State, which have their own experimental research programs, something we never got off the ground.

Speaking only about myself and not my teammates, while I was at MetaMed I was deeply *embarrassed* to be a layman in the biomedical field, and I felt like “why would an expert ever want to work with a layman like me?” So I was far too reluctant to reach out to prominent biologists and doctors. I now know that *experts work with laymen all the time*, especially when that layman brings strategic vision, funding, and logistical/operational manpower, and listens to the expert with genuine curiosity. Laymen are valuable — just ask Mary Lasker! I really wish I’d understood this at the time.

People overestimate progress in the short run and underestimate it in the long run. “Biohackers” and “citizen science” and “N=1 experimentation” have been around for a while, but they haven’t, I think, gotten very far along towards the ultimate impact they’re likely to have in the future. Naively, that can look a lot like “a few people tried that and it didn’t seem to go anywhere” when the situation is actually “the big break is still ahead of us.”

Rest Days vs Recovery Days

Based on a comment I made on this EA Forum Post on Burnout.

Related links: Sabbath hard and go home, Bring Back the Sabbath

---

That comment I made generated more positive feedback than usual (in that people seemed to find it helpful to read and found themselves thinking about it months after reading it), so I'm elevating it to a LW post of its own. Consider this an update to the original comment.

Like Ben Hoffman, I stumbled upon and rediscovered the Sabbath (although my implementation seems different from both Ben and Zvi). I was experiencing burnout at CFAR, and while I wasn't able to escape the effects entirely, I found some refuge in the following distinction between Rest Days and Recovery Days.

Recovery Days

A **Recovery Day** is where you're so tired or under-resourced that you can't do much of anything with yourself other than: stay in bed / sleep a lot, binge on Netflix or video games, stay in your room all day, play with your phone, use social media, and feel unmotivated to do much except easy, stimulating, and/or mind-numbing things. This is a Recovery Day and does not count as a Rest Day, but it is fine to take the time for them. However you aren't going to be refreshed from them. In order to really refresh, you need to take another day that counts as a Rest Day.

Another way a person might take time off is to do things that are *like work* but easier. Video games are a prime example. I play a lot of video games that involve optimizing systems, and I find these really motivating and fun. But I notice that this is a kind of "work"—my mind is trying to solve problems and implement solutions. The difference is that because it's easy and doable, I get addicted to them, and it's a way for me to escape the "real" problems at work, which tend to be harder to solve. This also doesn't count as Resting.

Rest Days

Rest Days are days where I have enough energy and resources that I feel motivated and able to get out and about. (One way I can tell I have energy is that sometimes I spontaneously feel like cooking, a rare occurrence.) On a Rest Day, your prime directive is to just "follow your gut" for the entire day and just do "what you feel like doing" in the moment.

There can be no obligations on a Rest Day. No scheduled calls or meetings. No promises to show up to a party. You can go to the party if you actually feel like going to the party, but you won't be able to know until last-minute. You cannot be "on-call" for anything. No one should depend on you unless it's someone you actively like being depended on for things, like a person you care about.

There can be exceptions to these, but I like to make Rest Days "sacred"—aka protected from influences like work pressure, social pressure, pressure from society, incentive gradients created by video games and my phone, incentive gradients created by money, the pressure to be different or better, the pressure to achieve, the pressure to always be going somewhere else, the pressure to "always be closing."

Rest Days are for being in the Now. The Now needs to be protected from influences from both the past (obligations) and the future (anxieties).

Rest Days will actually refresh and reset you. Unfortunately, many people do not know how to take Rest Days. They instead use weekends and vacation days as Recovery Days or days where their mind is still in "working" mode. But Recovery Days alone are not sufficient for refreshing your energy levels and motivation. You risk burnout if you consistently fail to get any true Rest over a long period of time.

Things my gut wants to do on Rest Days:

- be in the present moment
- meditate (in a natural, spontaneous way)
- eat tasty things
- have a picnic in a park / take walks / enjoy nature
- chill at a cafe I like
- go to a museum or an aquarium
- draw, dance, sing, improvise poems
- read a book, listen to music
- cook
- spend meaningful social time with friends or family
- useful, engaging errands or home-improvement stuff (not because I have to, because I want to)

Things my gut rarely wants to do on Rest Days:

- spend a lot of time on Facebook or social media
- binge TV
- play video games
- be in front of a screen in general
- do my job / work stuff
- lie in bed for hours
- eat microwaved food or junk food
- go to social functions, networking events, or any social event where I feel like I "should" go but don't really feel like going
- do anything with an addictive quality

Bottom-Up Implementation

My implementation of Rest Days / Sabbaths is very bottom-up. I pay attention to the sensations and signals from my stomach and use them as my guide for what to do and what not to do. It's basically using Focusing to play a game of warmer-colder on various actions I could take.

E.g.: I use this method all the time for deciding what to eat. I go through a list of possible foods I could eat, and I check each one by placing the image or felt sense of the food "next to" my stomach. The responses are usually like, "nah" or "not quite but closer" or "yes that." And if I check them against my mouth instead, the answers are sometimes different. My stomach tends to want "real food" (filling, satisfying, variety of nutrients) whereas my mouth will wants things based on their flavor (sweet, spicy, familiar, etc.).

I use the same method to decide what I want to do: go to the park? do some sketching? hang out with friends?

This kind of decision-making process doesn't work as well for complicated things. I'm not going to try to buy a house this way. Or plan a party. Or do any work. But it's a great way to know how to spend a Rest Day.

Top-Down Implementation

Another totally valid way to implement Rest Days is a top-down method, where you pre-determine some rules and guidelines for yourself.

Zvi has a set of simple rules he outlined in his post:

> Start here. Adjust as needed.
>
> Light candles before sundown Friday to begin.
>
> No outside inputs except in person.
>
> No choices impacting post-Sabbath.
>
> Light and extinguish no fires. Do no work or business. Spend no money.
>
> Only preselected and spontaneously motivated actions are allowed. No browsing. No lists.
>
> Light another candle after sundown Saturday to end.

Some other pick-and-choose options for rules that I think would work for many people:

- No social media, email, news, or mindless phone games
- No obligations or prior commitments (only optional activities)
- No driving
- No online shopping / no buying anything that costs more than $50 unless it's a spontaneous gift for someone else
- (If EA) Try to make this day one where you permit yourself to seek out warm fuzzies and set aside questions of utility / doing the most good
- No planning for the future
- Give yourself "Get Out of Jail Free" cards for things like social obligations, community drama/issues, work-related problems

Fair warning #1: If you go overboard on the rules, you may never discover what *truly resting* is like for you, as I believe it is different for each person AND I don't think you can know what *resting* is for you without checking in that exact moment in time. Resting is about NOW. Trying to "get your future self to rest" by outlining a bunch of rules may cause you to miss some important things about what you're really wanting in the moment.

True Rest is one where, in the moment, you do what you want to do and don't do what you don't want to do. That's it.

Fair warning #2: If you give yourself too much room to maneuver, you may end up slipping back into old habits and just treating Rest Days like any other day. Maybe you say to yourself, well I *really* actually feel like doing this work right now. So you do some work. And then the next time, it happens again. And again. Until it spirals into becoming normal to work on Rest Days—to pick up work calls, to schedule meetings, to check email, etc.

Rest Days deserve sacred levels of protection. Otherwise you will just lose them.

I don't expect anyone to be able to have perfect Rest Days.

I still check email and Facebook on my Rest Days, just less often. If a work emergency came up, I'd probably get pulled in.

Fine.

But I think it makes a significant difference even just to a) hold it as your intention to Rest for the day and b) let others know that this is important to you and that they would be impinging by making requests of you on a Rest Day. This is *your time*. You are allowed to set boundaries on your time, your attention, and your energy.

Even if you can't pull it off every week, it seems good to at least try for once a month. Twelve days out of the year are for you. And hopefully it's closer to fifty.

The Sabbath was trivial to enforce when everyone was doing it. We've more or less lost that as a shared norm. As such, you will be fighting an upstream battle to hold onto your sacred Rest Days. This is unfortunate.

But it is worth it.

In my culture, you are allowed to stand up for your ability to Rest. To say "Fuck you" to outside forces trying to take that away from you. To get angry, to dig your heels in, to be stubborn, to get indignant. To say no. You are allowed to protect the sacredness of your Rest Day.

Society has mostly given up on sacred Rest Days. The least I can do is make it openly permissible and defensible for you to protect your ability to have Rest Days. I hope we can do the same for each other.

You Get About Five Words

Cross posted from the EA Forum.

Epistemic Status: all numbers are made up and/or sketchily sourced. Post errs on the side of simplistic poetry – take seriously but not literally.

If you want to coordinate with one person on a thing about something nuanced, you can spend as much time as you want talking to them – answering questions in realtime, addressing confusions as you notice them. You can trust them to go off and attempt complex tasks without as much oversight, and you can decide to change your collective plans quickly and nimbly.

You probably speak at around 100 words per minute. That's 6,000 words per hour. If you talk for 3 hours a day, every workday for a year, you can communicate 4.3 million words worth of nuance.

You can have a real conversation with up to 4 people.

(Last year the small organization I work at considered hiring a 5th person. It turned out to be very costly and we decided to wait, and I think the reasons were related to this phenomenon)

If you want to coordinate on something nuanced with, say, 10 people, you realistically can ask them to read a couple books worth of words. A book is maybe 50,000 words, so you have maybe 200,000 words worth of nuance.

Alternately, you can monologue at people, scaling a conversation past the point where people realistically can ask questions. Either way, you need to hope that your books or your monologues happen to address the particular confusions your 10 teammates have.

If you want to coordinate with 100 people, you can *ask* them to read a few books, but chances are they won't. They might all read a few books worth of stuff, but they won't all have read the same books. The information that they can be coordinated around is more like "several blogposts." If you're trying to coordinate *nerds*, maybe those blogposts add up to one book because nerds like to read.

If you want to coordinate 1,000 people... you realistically get one blogpost, or maybe one blogpost worth of jargon that's hopefully self-explanatory enough to be useful.

If you want to coordinate thousands of people...

You have about five words.

This has ramifications on how complicated a coordinated effort you can attempt.

What if you *need* all that nuance *and* to coordinate thousands of people? What would it look like if the world was filled with complicated problems that required lots of people to solve?

I guess it'd look like this one.

The Amish, and Strategic Norms around Technology

I was reading Legal Systems Very Different From Ours by David Friedman. The chapter on the Amish made a couple interesting claims, which changed my conception of that culture (although I'm not very confident that the Amish would endorse these claims as fair descriptions).

Strategic Norms Around Technology

The Amish relationship to technology is not "stick to technology from the 1800s", but rather "carefully think about how technology will affect your culture, and only include technology that does what you want."

So, electric heaters are fine. *Central* heating in a building is not. This is because if there's a space-heater in the living room, this encourages the family to congregate together. Whereas if everyone has heating in their room, they're more likely to spend time apart from each other.

Some communities allow tractors, but only if they don't have rubber tires. This makes them good for tilling fields but bad for driving around.

Cars and telephones are particularly important not to allow, because easy transportation and communication creates a slippery slope to full-connection to the outside world. And a lot of the Amish lifestyle depends on cutting themselves off from the various pressures and incentives present in the rest of the world.

Some Amish communities allow people to borrow telephones or cars from non-Amish neighbors. I might have considered this hypocritical. But in the context of "strategic norms of technology", it need not be. The important bit is to *add friction* to transportation and communication.

Competitive Dictatorship

Officially, most Amish congregations operate via something-like-consensus (I'm not sure I understood this). But Friedman's claim is that in practice, most people tend to go with what the local bishop says. This makes a bishop something like a dictator.

But, there are lots of Amish communities, and if you don't like the direction a bishop is pushing people in, or how they are resolving disputes, you can leave. There is a spectrum of communities ranging in how strict they are about about various rules, and they make decisions mostly independently.

So there is not only strategic norms around technology, but a fairly interesting, semi-systematic exploration of those norms.

Other Applications

I wouldn't want to be Amish-in-particular, but the setup here is very interesting to me.

I know some people who went to MAPLE, a monastery program. While there, there were limits on technology that meant, after 9pm, you basically had two choices: read, or go to bed. The choices were strongly reinforced by the social and physical environment. And this made it much easier to make choices they endorsed.

Contrast this with my current house, where a) you face basically infinite choices about to spend your time, and b) in practice, the nightly choices often end up being something like "stay up till 1am playing minecraft with housemates" or "stay up till 2am playing minecraft with housemates."

I'm interested in the question "okay, so... my goals are not the Amish goals. But, what *are* my goals exactly, and is there enough consensus around particular goals to make valid choices around norms and technology other than 'anything goes?'"

There are issues you face that make this hard, though:

Competition with the Outside World – The Amish system works because it cuts itself off from the outside world, and its most important technological choices directly cause that. Your business can't get outcompeted by someone else who opens up their shop on Sundays because there is nobody who opens their shop on Sundays.

You also might have goals that directly involve the outside world.

(The Amish also have good relationships with the government such that they can get away with implementing their own legal systems and get exceptions for things like school-laws. If you want to do something on their scale, you *both* would need to not attract the ire of the government, and be good enough at rolling your own legal system to not screw things up and drive people away)

Lack of Mid-Scale-Coordination – I've tried to implement 10pm bedtimes. It fails, horribly, because I frequently attend events that last till midnight or later. *Everyone* could shift their entire sleep schedule forward, maybe. But also...

People Are Different – *Some* of people's needs are cultural. But some are biological, and some needs are maybe due to environmental factors that happened over decades and can't be changed on a dime.

Some people do better with rules and structure. Some people flourish more with flexibility. Some people need rules and structure but *different* rules and structure than other people.

This all makes it fairly hard to coordinate on norms.

Contenders for Change

Given the above, I think it makes most sense to:

- Look for opportunities explore norms and technology-use at the level of individuals, households, and small organizations (these seem like natural clusters with small numbers of stakeholders, where you can either get consensus or have a dictator).

- While doing so, choose norms that are locally stable, that don't require additional cooperation outside yourself, your household or your org.

For example, I could imagine an entire household trying out a rule, like "the household internet turns off at 10pm", or "all the lights turn reddish at night so it's easier to get to sleep"

Dependability

I have become very, very interested in developing a skill that I call **Dependability**.

I believe the skill exists on a spectrum, and you can have less or more of it. It's not a binary where you either have it or you don't.

I believe this skill can be trained on purpose.

I will briefly describe each attribute that makes up the overall skill.

(All the examples are made up. Also, assume that the examples are only talking about *endorsed* actions and goals.)

Trying

To have a vision of a skill or a desirable end state and then be able to strive with deliberate effort towards making that vision a reality.

Ex1: I see a dance routine on YouTube that I think would be awesome if I could perform myself. I've never done anything like this before, and I'm somewhat self-conscious or skeptical of how likely I am to succeed. Regardless, I take concrete steps towards learning it (study the video, practice the moves that I see, repeat for many days until I've achieved competency at performing the dance). There is some possibility I fail for whatever reason, but this doesn't stop me from giving it my full effort for at least a week.

Ex2: There's a job that I really want. I'm unclear about what steps I need to take to acquire the job, and I'm not sure I'm qualified. I research what kinds of skills and traits are desirable in the job by asking people, Googling, and looking through applications. (I am more encouraged than not by this initial research.) I sign up for workshops and classes that will give me relevant training. I read books. I practice in my free time. I make useful connections / network. I build whatever reputational capital seems useful via blogging, social media, in-person meetings, running events. I apply for the job. If I fail, I figure out what needs work, fix it, and try again until I obtain the position.

Commitment

To form an intention to do something (generally on a longer time scale), be able to say it out loud to someone else, and then be certain it will happen one way or another, barring extreme circumstance.

Ex1: I commit myself via marriage to another person and promise that I will try everything to make the relationship work before giving up on it. I say it out loud as a vow to the other person in a marriage ceremony, in front of a bunch of people. Then I proceed to actually attempt to get as close to 100% chance of creating a permanent relationship situation with this person, using all the tools at my disposal.

Ex2: I tell someone that I will be there for them in times of emergency or distress, if they ask. I tell them I will make it a priority to me, over whatever else is going on in my life. A year or two later (possibly with very little contact with this person otherwise), they call me and ask for my help. I put everything aside and create a plan to make my way to them and provide my assistance.

Follow-through

To finish projects that you start, to not give up prematurely, to not lose the wind in your sails out of boredom, lack of short-term incentive or immediate reward, lack of encouragement, or feelings of uncertainty and fear.

Here's an example of what it looks like to NOT have follow-through: I want to write a novel, but every time I start, I lose interest or momentum after initial drafting and planning. Maybe I manage to build the world, create characters, plan out a plot, but then I get to the actual writing, and I fail to write more than a few chapters. Or maybe I loosen the requirements and decide I don't need to plan everything out in advance, and I just start writing, but I lose steam midway through. I know in my heart that I will never be able to finish it (at least, without some drastic change).

Having follow-through means having the ability to finish the novel to completion. It is somehow missing from the person I've described above.

Reliability

To do what you say you'll do (on a lesser scale than with commitment); to be where you say you'll be, when you say you'll be there; to cooperate proactively, consistently, and predictably with others when you've established a cooperative group dynamic.

This can also be summed up as: If you set an expectation in someone else, you don't do something that would dramatically fail to meet their expectation. You either do the thing or you communicate about it.

Examples:

If someone is expecting to meet me at a time and place, I show up at the time and place. If there are delays, I let them know ahead of time. I don't ever fail to show up AND not tell them in advance AND not explain afterwards (this would count as dramatically failing to meet an expectation).

If someone asks me to complete a task within the month, and months later, I have both failed to do the task AND I have become incommunicado, this counts as dramatically failing to meet an expectation.

Note that it doesn't actually matter if they feel upset by your failure to meet an expectation. They might be totally fine with it. But I still would not have the skill of reliability, by this definition.

The skill also includes an ability to "plug into" teams and cooperative situations readily. If you are on a team, you are relatively easy to work with. You communicate clearly and proactively. You take responsibility for the tasks that are yours.

Focused attention

To be able set an intention and then keep your attention on something for a set amount of time (maybe about up to 20 minutes).

Ex1: If someone I care about is speaking to me and what they're saying is important to them, even if it isn't that important to me, I am able to pay attention, hear their words, and not get lost in my own thoughts such that I can no longer attend to their words.

Ex2: If I am trying to complete a <20-min task, I do not get distracted by other thoughts. I do not follow every impulse or urge to check Facebook or play a game or get food, such that I cannot complete the task. I'm able to stay focused long enough to finish the task.

Being with what is

To not flinch away from what is difficult, aversive, or painful. To be able to make space for sensations and emotions and thoughts, even if unpleasant. To be able to hold them in your mind without following an automatic reaction to move away or escape.

Ex1: If I am trying to introspect on myself, and I encounter ughy, aversive, or uncomfortable feelings, thoughts, or realizations, I am able to make space for that in my mind and stay with them. (This probably involves distancing myself somewhat from them so that they're not overwhelming.)

Ex2: If someone expresses a loud, big, "negative" emotion (anger, fear, sadness, pain), I don't panic or freeze or dissociate. I can stay calm, embodied, and grounded. And then I stay open to their emotional state and not assume it means something bad about me ("They hate me!" "I'm doing something wrong!" "They don't want me around!"). I'm not overwhelmed by anxieties or stories about what their emotion means, which might cause me to go away or stop caring about them. I instead make room in myself for my feelings and their feelings so that they can both exist. I maintain an open curiosity about them.

More thoughts on Dependability

I claim that all these skills are tied together and related in some important way, and so I bundle them all under the word Dependability. Although I do not myself understand exactly and precisely how they're related.

My sense is that the smaller-scale skills (e.g. focused attention, which occurs on a moment-to-moment scale) add to your ability to achieve the larger-scale skills (e.g. commitment, which occurs on a month-to-month scale).

If I had to point to the *core* of the Dependability skill and what the foundation of it is, it is based on two things: the ability to set an intention and the ability to stay with what is. And all the above skills apply these two things in some way.

In general, people seem able to set intentions, but the “staying” is the tricky part. Most people I’ve encountered have some of the Dependability skill, to some extent. But the skill is on a spectrum, and I’d grade most people as “middling.”

I think I’m personally much worse at setting intentions than average. In certain domains (emotions, realizations), I’m above average at staying with what is. In other domains (failure, setbacks, physical discomfort), I’m much, much worse at staying with what is.

I suspect children are not born with the overall skill. They develop it over time. The marshmallow test seems to assess part of the skill in some way?

My stereotype of a typical high school or college kid (relative to an adult) is terrible at the overall skill, and especially reliability. I was a prime example. You couldn’t rely on me for anything, and I was really bad at communicating the ways in which I was unreliable. So I just fell through on people a lot, especially people with authority over me. I would make excuses, ask for extensions and exceptions, and drop the ball on things.

Over time, I learned to do that way less. I’ve drastically improved in reliability, which was helped by having a better self-model, learning my limitations, and then setting expectations more appropriately. I’ve also just obtained more object-level skills such that I can actually do more things. I’ve learned to extend my circle of caring to beyond just myself and my needs, so I can care about the group and its needs.

The other skills, however, I am still quite bad at. Some of them I’m completely incapable of (commitment, follow-through).

How do you train Dependability?

I personally feel crippled without the skill. Like I will never achieve my most important goals without it. And also, I feel particularly disabled in gaining the skill, because of how I reacted to childhood trauma. My way of being, so far, has completely avoided making commitments, trying, and having follow-through. I’ve found workarounds for all those things such that I’ve lived my life without having to do them. And I got by just fine, but I won’t be able to achieve many of my goals this way.

(It’s a blessing and a curse that an intelligent, precocious person can get by without the *trying* skill, but here we are...)

Fortunately for me, I currently believe the skill is trainable with deliberate practice. Possibly better in combination with introspective, therapeutic work.

I don’t know what kind of training would work for others, but for myself, I’ve found one plausible way to train the skill deliberately.

I spent a week at a place called MAPLE, aka the Monastic Academy for the Preservation of Life on Earth. The people I met there exhibited above average skill in Dependability, and I was notably surprised by it. I was so surprised by it that I’ve spent a lot of time thinking about MAPLE and talking to people about it. And now I’ll be spending a month there as a trial resident, starting in April.

But this post isn't where I talk about MAPLE. I mention it primarily as a hint that maybe this skill is attainable through deliberate practice.

It kind of makes sense that very deliberate, regular meditation could contribute to the skill. Because maybe the micro-skill (setting lots of tiny intentions, being with what is on a moment-to-moment basis) contributes to the macro-skill (setting large intentions, staying with what is on a larger scale).

The monastic lifestyle also includes being tasked with all kinds of somewhat aversive things (cleaning bathrooms, managing people, being responsible for things you've never been responsible for before). You join the team and are expected to contribute in whatever ways are needed to maintain and run the monastery. And it is supposed to be hard, but you are training even then.

It seems possible that this month at MAPLE, I will set more deliberate intentions than I have collectively in my life until then. Which tells you just how little I've done things *on purpose, deliberately, and with intention* in my life. The process of how that got broken in me is probably another story for another time.

But basically, I expect to do a bunch of repetitions of training Dependability on a second-to-second level. And I will be doing this not just during meditation but also during daily work. I will also likely spend a lot of time introspecting and trying to gain insight into my blocks around Dependability. I hope to see at least a little movement in this area in the next month but may need to spend a longer period of time at MAPLE to fully develop the skill. (I noticed that residents who'd been at MAPLE for multiple years had more of the skill than those who had been there for less time.)

[Note: The following section might trigger people who are scrupulous in a particular way. I want to make clear that I'm not speaking from a place of obligation or shouldy-ness or fear of being a bad or unworthy person or self-judgment. I don't feel shame or guilt about not having Dependability. I'm speaking from a place of actively wanting to grow and feeling excited about the possibility of attaining something important to me. And I hope the same for other people, that they will be motivated *towards* having nice things. Dependability seems like a nice thing to have, but I'm not into judging people (or myself) about it.]

Not having Dependability is a major bottleneck for me. My ultimate goal is to live a life of arete, or excellence in all things. And an especially important part of that for me is living a virtuous life.

I believe that without Dependability, I will not be able to live a virtuous life: Be the kind of person who makes correct but difficult choices. Be the kind of person who is reliably there for her friends and family. Be the kind of person who can become part of or contribute to something bigger than herself. Be the kind of person who wouldn't sell out humanity for money, fame, power, convenience, security, legacy. Be the kind of person who doesn't lie to herself about "being a good person" who "does things for the sake of progress or for the good of others"—when in truth the underlying behaviors, cruxes, and motives have little to do with the rationalizations.

I consider it my duty as a human being to develop into a virtuous person, rather than just any kind of person. And I believe Dependability is an important feature of a virtuous person.

I notice I don’t meet my personal criteria for a virtuous person as of yet, and Dependability seems like a major missing piece.

Privacy

Follow-up to: Blackmail

[Note on Compass Rose response: This is *not* a response to the recent Compass Rose response, it was written before that, but with my post on Hacker News I need to get this out now. It *has* been edited in light of what was said. His first section is a new counter-argument against a particular point that I made – it is interesting, and I have a response but it is beyond scope here. It does not fall into either main category, because it is addressing a particular argument of mine rather than being a general argument for blackmail. The second counter-argument is a form of #1 below, combined with #2, #3 and #4 (they do tend to go together) so it *is* addressed somewhat below, especially the difference between 'information tends to be good' and 'information chosen, engineered and shared so to be maximally harmful tends to be bad.' My model and Ben's of practical results also greatly differ. We intend to hash all this out in detail in conversations, and I hope to have a write-up at some point. Anyway, on to the post at hand.]

There are two main categories of objection to my explicit thesis that blackmail should remain illegal.

Today we will not address what I consider the more challenging category. Claims that while blackmail is bad, making it illegal does not improve matters. Mainly because we can't or won't enforce laws, so it is unclear what the point is. Or costs of enforcement exceed benefits.

The category I address here claims blackmail is good. We want more.

Key arguments in this category:

1. Information is good.*
2. Blackmail reveals bad behavior.
3. Blackmail provides incentive to uncover bad behavior.
4. Blackmail provides a disincentive to bad behavior.
5. Only bad, rich or elite people are vulnerable to blackmail.
6. We should strongly enforce all norms on everyone, without context dependence not explicitly written into the norm, and fix or discard any norms we don't want to enforce in this way.

A key assumption is that blackmail mostly targets *existing true bad behavior.* I do not think this is true. For true *or* bad *or* for existing. For details, see the previous post.

Such arguments also centrally argue against *privacy.* Blackmail advocates often claim privacy is unnecessary or even toxic.

It's one thing to give up on privacy in practice, for yourself, in the age of Facebook. I get that. It's another to argue that *privacy is bad.* That *it is bad to not reveal all the information you know.* Including about yourself.

This radical universal transparency position, perhaps even *assumption,* comes up quite a lot recently. Those advocating it act as if those opposed carry the burden of proof.

No. Privacy is good.

A reasonable life, a good life, *requires* privacy.

I

We need a realm *shielded from signaling and judgment.* A place where what we do *does not change what everyone thinks about us, or get us rewarded and punished.* Where others don't judge what we do based on the assumption that we are choosing what we do knowing that others will judge us based on what we do. Where we are free from others' Bayesian updates and those of computers, from what is correlated with what, with how things look. A place to play. A place to experiment. To unwind. To celebrate. To learn. To vent. To be afraid. To mourn. To worry. To be yourself. To *be real.*

We need people there with us *who won't judge us.* Who won't *use information against us.*

We need having such trust to not risk our ruin. We need to minimize how much we wonder, if someone's goal is to get information to use against us. Or what price would tempt them to do that.

Friends. We desperately need real friends.

II

Norms are not laws.

Life is full of trade-offs and necessary unpleasant actions that violate norms. This is not a fixable bug. Context is important for both enforcement and intelligent or useful action.

Even if we could fully enforce norms in principle, different groups have different such norms and each group's/person's norms are self-contradictory. Hard decisions mean violating norms and are common in the best of times.

A complete transformation of our norms and norm principles, beyond anything I can think of in a healthy historical society, would be required to even attempt full non-contextual strong enforcement of all remaining norms. It is unclear how one would avoid a total loss of freedom, or a total loss of reasonable action, productivity and survival, in such a context. Police states and cults and thought police and similar ideas have been tried and have definitely not improved this outlook.

What we do for fun. What we do to make money. What we do to stay sane. What we do for our friends and our families. What maintains order and civilization. *What must be done.*

Necessary actions are often the very things others wouldn't like, or couldn't handle… *if revealed in full, with context simplified to what gut reactions can handle*.

Or worse, *with context chosen to have the maximally negative gut reactions.*

There are also known dilemmas where *any action taken* would be a norm violation of a sacred value. And lots of values that *claim* to be sacred, because every value wants to be sacred, but which we know *we must treat as not sacred* when making real decisions with real consequences.

Or in many contexts, justifying our actions would require revealing massive amounts of private information that would then cause further harm (and which people very much do not have the time to properly absorb and consider). Meanwhile, you're taking about the bad-sounding thing, which digs your hole deeper.

We all must do these necessary things. These often violate both norms and formal laws. Explaining them often requires sharing other things we dare not share.

I wish everyone a past and future Happy Petrov Day

Part of the job of making sausage is to allow others not to see it. We still get reliably disgusted when we see it.

We *constantly* must claim 'everything is going to be all right' or 'everything is OK.' That's *never* true. Ever.

In these, and in many other ways, we live in an unusually hypocritical time. A time when people need be far more afraid both to not be hypocritical, and of their hypocrisy being revealed.

We are a nation of men, not of laws.

But these problems, while improved, wouldn't go away in a better or less hypocritical time. Norms are not a system that can have full well-specified context dependence and be universally enforced. That's not how norms work.

III

Life requires privacy so we can *not reveal the exact extent of our resources.*

If others know exactly what resources we have, they can and will take all of them. The tax man who knows what you *can* pay, what you *would* pay, already knows what you *will* pay. For government taxes, and for other types of taxes.

This is not only about payments in money. It is also about time, and emotion, and creativity, and everything else.

Many things in life claim to be sacred. Each claims all known available resources. Each claims we are blameworthy for any resources we hold back. If we hold nothing back, we have nothing.

That which is fully observed cannot be one's slack. Once all constraints are known, they bind.

Slack requires privacy. Life requires slack.

The includes our *decision making process.*

If it is known how we respond to any given action, others find best responses. They will respond to incentives. They exploit *exactly* the amount we won't retaliate against. They feel safe.

We seethe and despair. We have no choices. No agency. No slack.

It is a key protection that one *might* fight back, perhaps massively out of proportion, if others went after us. To any extent.

It is a key protection that one *might* do something good, if others helped you. Rather than others knowing *exactly* what things will cause you to do good things, and which will not.

It is central that one *react when others are gaming the system.*

Sometimes that system is you.

World peace, and doing anything at all that interacts with others, depends upon both *strategic confidence* in some places, and *strategic ambiguity* in others. We need to choose carefully where to use which.

Having all your actions fully predictable and all your information known isn't Playing in Hard Mode. That's Impossible Mode.

I now give specific responses to the six claims above. This mostly summarizes from the previous post.

1. Information, by default, is probably good. But this is a tenancy, not a law of physics. As discussed last time, information *engineered to be locally harmful* probably is net harmful. Keep this distinct from incentive effects on bad behavior, which is argument number 4.
2. Most 'bad' behavior will be a justification for scapegoating, involving levels of bad behavior that are common. Since such bad behavior is rarely made common knowledge, and allowing it to become common knowledge *is often considered far worse behavior than the original action,* making it common knowledge forces oversize reaction and punishment. What people are punishing is *that you are the type of person who lets this type of information become common knowledge about you.* Thus you are not a good ally. In a world like ours, where all are anticipating future reactions by others anticipating future reactions, this can be devastating.
3. Blackmail does provide incentive to investigate to find bad behavior. But if found, it also provides incentive to make sure it is never discovered. And what is extracted from the target is often further bad behavior, largely because…
4. Blackmail also provides an incentive to *engineer or provoke* bad behavior, and to maximize the damage that would result from revelation of that behavior. The incentives promoting more bad behavior likely *are stronger* than the ones discouraging it. I argue in the last piece that it is common *even now* for people to engineer blackmail material against others *and often also against themselves,* to allow it to be used as collateral and leverage. That a large part of job interviews is proving that you are vulnerable in these ways. That much bonding is about creating mutual blackmail material. And so on. This seems quite bad.
5. If any money one has can be extracted, then one will permanently be broke. This is a lot of my model of poverty traps – there are enough claiming-to-be-sacred things demanding resources that any resources get extracted, so no one tries to acquire resources or hold them for long. Consider what happens if people in such situations are allowed to borrow money. Even if you are (for any reason) sufficiently broke that you cannot pay money, you have much that you could be forced to say or do. Often this involves deep compromises of sacred values, of ethics and morals and truth and loyalty and friendship. It often involves being an ally of those you despise, and reinforcing that which is making your life a living hell, to get the pain to let up a little. Privacy, and the freedom from blackmail, are the only ways out.

6. A full exploration is beyond scope but section two above is a sketch.

* – I want to be very clear that *yes, information in general is good.* But that is a far cry from the radical claim that all and any information is good and sharing more of it is everywhere and always good.

How to Understand and Mitigate Risk

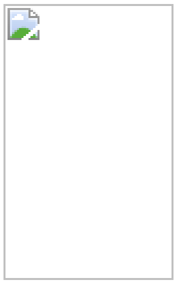
This post wouldn't be possible without support from the EA Hotel.

Epistemic Status: Fairly certain these distinctions are pointing at real things, less certain that the categories are exactly right. There's still things I don't know how to fit into this model, such as using Nash Equilibria as a strategy for adversarial environments.

Instrumental Status: Very confident that you'll get better outcomes if you start using these distinctions where previously you had less nuanced models of risk.

(larger image link)

Transparent Risks



Transparent risks are those risks that can be easily quantified and known, in advance. They're equivalent to the picture above, with a transparent bag where I can count the exact amount of marbles in each bag. If I'm also certain about how much each marble is worth, then I have a simple strategy for dealing with risks in this situation.

How to Mitigate Transparent Risks: Do the Math

The simple strategy for transparent risks like the one above is to do the math.

Expected Value

Expected value is a simple bit of probability theory that says you should multiply the likelihood of an event happening by the payoff to get your long run value over time. It's a simple way to figure out if the risk is worth the reward in any given situation. The best introduction I know to expected value is here.

Kelly Criterion

The Kelly criterion is helpful when losing your entire bankroll is worse than other outcomes. I don't fully understand it, but you should, and Zvi wrote a post in it here. (If someone would be willing to walk me through a few examples and show me where all the numbers in the equation come from, I'd be very grateful.)

Transparent Risks in Real Life

Drunk Driving

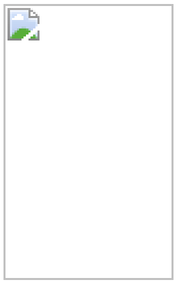
Driving drunk is a simple, well studied risk on which you can quickly find probabilities of crash, injury and death to yourself and others. By comparing these costs to the costs of cab fare (and the the time needed to get your car in the morning if you left it), you can make a relatively transparent and easy estimate whether it's worth driving at your Blood Alcohol Content level (spoiler alert, No if your BAC is anywhere near .08 on either side.) The same method can be used for any well-studied risks that exist within tight, slow changing bounds.

Commodity and Utility Markets

While most business opportunities are not transparent risks, an exception exists for commodities and utilities (in the sense mean't by Wardley Mapping). It's quite easy to research the cost of creating a rice farm, or a power plant, as well as get a tight bounded probability distribution for the expected price you can sell your rice or electricity at after making the initial investment. These markets are very mature and there's unlikely to be wild swings or unexpected innovations that significantly change the market. However, because these risks are transparent it also means that competition drives margins down. The winners are those which can squeeze a little extra margin through economies of scale or other monopoly effects like regulatory capture.

Edit: After being pointed to the data on commodities, I no longer lump them in with utilities as transparent risks and would call them more Knightian.

Opaque Risks



Opaque risks are those risks that can be easily quantified and unlikely to change, but which haven't already been quantified/aren't easy to quantify just by research. They're equivalent to the picture above, with an opaque bag that you know contains a static amount of a certain type of marble, but not the ratio of marbles to each other. As long as I'm sure that the bag contains only three types of marbles, and that the distribution is relatively static, a simple strategy for dealing with these risks emerges.

How to Mitigate Opaque Risks: Determine the Distribution

The simple strategy for opaque risks is to figure out the distribution. For instance, by pulling a few marbles at random out of the bag, you can over time become more and more sure about the distribution in the bag, at which point you're now dealing with transparent risks. The best resource I know of for techniques to determine the distribution of opaque risks is How to Measure Anything by Douglas Hubbard.

Sampling

Sampling involves repeatedly drawing from the distribution in order to get an idea of what the distribution is. In the picture above, it would involve simply reaching your hand in and pulling a few marbles out. The bigger your sample, the more sure you can be about the underlying distribution.

Modelling

Modelling involves breaking down the factors that create the distribution, into as transparent pieces as possible. The classic example from fermi estimation is how many piano tuners there are in Chicago - that number may be opaque to you, but the number of people in Chicago is relatively transparent, as is the percentage of people that own pianos, the likelihood that someone will want their piano tuned, and the amount of money that someone needs to make a business worthwhile. These more transparent factors can be used to estimate the opaque factor of piano tuners.

Opaque Risks in Real Life

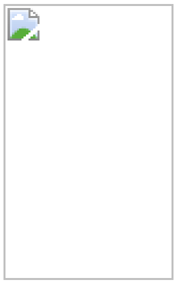
Choosing a Career You Don't Like

In the personal domain, opaque risks often take the form of very personal things that have never been measured because they're unique to you. As a career coach, I often saw people leaping forward into careers that were smart from a global perspective (likely to grow, good pay, etc) but ignored the more personal factors. The solution was a two tier sampling solution: Do a series of informational interviews for the top potential job titles and potential industries, and then for the top 1-3 careers/industries, see if you can do a form of job shadowing. This significantly helped cut down the risk by making an opaque choice much more transparent.

Building a Product Nobody Wants

In the business domain, solutions that are products(in Wardley Mapping terms) but are not yet commoditized often qualify as opaque risks. In this case, simply talking to customers, showing them a solution, and asking if they'll pay, can save a significant amount of time and expense before actually building the product. Material on "lean startup" is all about how to do efficient sampling in these situations.

Knightian Risks



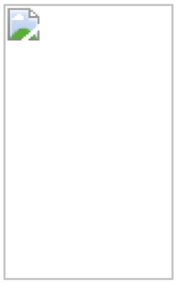
Knightian risks are those risks that exist in environments with distributions that are actively resistant to the methods used with opaque risks. There are three types of Knightian Risks: Black Swans, Dynamic Environments, and Adversarial Environments.

A good portion of "actually trying to get things done in the real world" involves working with Knightian risks, and so most of the rest of this essay will focus on breaking them own into their various types, and talking about the various solutions to them.

Milan Griffes has written about Knightian risks in an EA context on the EA forum, calling them "cluelessness".

Types of Knightian Risks

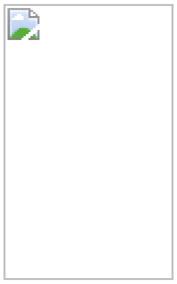
Black Swans



A black swan risk is an unlikely, but very negative event that can occur in the game you choose to play.

In the example above, you could do a significant amount of sampling without ever pulling the dynamite. However, this is quite likely a game you would want to avoid given the presence of the dynamite in the bag. You're likely to severely overestimate the expected value of any given opportunity, and then be wiped out by a single black swan. Modelling isn't useful because very unlikely events probably have causes that don't enter into your model, and it's impossible to know you're missing them because your model will appear to be working accurately (until the black swan hits). A great resource for learning about Black Swans is the eponymous Black Swan, by Nassim Taleb.

Dynamic Environments

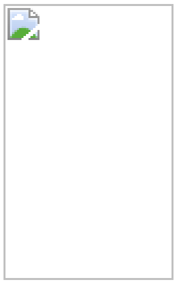


When your risks are changing faster than you can sample or model them, you're in a dynamic environment. This is a function of how big the underlying population size is, how good you are at sampling/modelling, and how quickly the distribution is changing.

A traditional sampling strategy as described above involves first sampling, finding out your risks in different situations, then finally "choosing your game" by making a decision based on your sample. However, when the underlying distribution is changing rapidly, this strategy is rendered moot as the information your decision was based on quickly becomes outdated. The same argument applies to a modelling strategy as well.

There's not a great resource I know of to really grok dynamic environments, but an ok resource is Thinking in Systems by Donella Meadows (great book, but only ok for grokking the inability to model dynamic environments).

Adversarial Environments

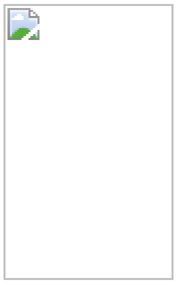


When your environment is actively (or passively) working to block your attempts to understand it and mitigate risks, you're in an adversarial environment.

Markets are a typical example of an Adversarial Environment, as are most other zero sum games with intelligent opponents. They'll be actively working to change the game so that you lose, and any change in your strategy will change their strategy as well.

Ways to Mitigate Knightian Risks

Antifragility



Antifragility is a term coined by Nassim Taleb to describe systems that gain from disorder. If you think of the games described above as being composed of distributions, and then payoff rules that describe how you react to this distributions, anti-fragility is a look at how to create flexible payoff rules that can handle Knightian risks. Taleb has an excellent book on anti-fragility that I recommend if you'd like to learn more.

In terms of the "marbles in a bag" metaphor, antifragility is a strategy where pulling out marbles that hurt you makes sure you get less and less hurt over time.

- *Optionality*

Optionality is a heuristic that says you should choose those options which allow you to take more options in the future. The idea here is to choose policies that lower you're intertia and switching costs between strategies. Avoiding huge bets and long time horizons that can make our break you, while developing agile and nimble processes that can quickly change. This is the principle from which all other anti-fragile principles are generated.

This helps with black swans by allowing you to quickly change strategies when your old strategy is rendered moot by a black swan. It helps with dynamic environments by allowing your strategy to change as quickly as the distribution does. It helps with adversarial environments by giving you more moves to use against changing opponents.

Going with the bag of marbles example, imagine there are multiple bags of marbles, and the distributions are changing over time. Originally, it costs quite a lot to switch between bags. The optionality strategy says you should be focused on lowering the cost of switching between bags over time.

- *Hormesis*

Hormesis is a heuristic that says that when negative outcomes befall you, you should work to make that class of outcomes less likely to hurt you in the future. When something makes you weak temporarily, you should ultimately use that to make yourself stronger in the long run.

This helps with Black Swans by gradually building up resistance to certain classes of black swans BEFORE they hit you. It helps with rapidly changing distributions by continually adapting to the underlying changes with hormetic responses.

In the bag of marbles example, imagine that at the start pulling a red marble was worth -$10. Every time you pulled a red marble, you worked to reduce that harm of red things by 1/10. This would mean that in an environment with lots of red marbles, you would quickly become immune to them. It would also mean that if you eventually did pull out that stick of dynamic, your general ability to handle red things would mean that it would hurt you less.

(I get that the above example is a bit silly, but the general pattern of immunity to small events helping you with immunity to black swans in the same class is quite common).

- *Evolution*

The evolution heuristic says that you should constantly be creating multiple variations on your current strategies, and keeping those that avoid negative consequences over time. Just like biological evolution, you're looking to find strategies that are very good at survival. Of course, you should be careful about calling up blind idiot gods, and be cautious about being tempted to optimize gains instead of minimize downside risk (as it should be used).

This helps with black swans in a number of ways. Firstly, by diversifying your strategies, it's unlikely that all of them will be hit by black swans. Secondly, it has an effect similar to hormesis in which immunity to small effects can build up immunity to black swans in the same class. Finally, by having strategies that outlive several black swans, you develop general survival characteristics that help against black swans in general. It helps with dynamic environments by having several strategies, some of which will hopefully be favorable to the environmental changes.

- *The Barbell Strategy*

The barbell strategy refers to a strategy of splitting your activities between those that are very safe, with low downside, and those that are very risky, with high upside. Previously, Benquo has argued against the barbell strategy, arguing that there is no such thing a riskless strategy. I agree with this general idea, but think that the framework I've provided in this post gives a clearer way to talk about what Nassim means: Split your activities between transparent risks with low downsides, and Knightian risks with high upsides.

The transparent risks obviously aren't riskless (that's why they're called risk), but they behave relatively predictably over long time scales. When they DON'T behave predictably is when there's black swans, or an equilibrium is broken such that a relatively stable environment becomes an environment of rapid change. That's exactly when the transparent risks with high upside tend to perform well (because they're designed to take advantage of these situations). That's also why this strategy is great for handling black swans and dynamic environments. It's less effective at handling adversarial environments, unless there's local incentives in the adversarial environment to think more short term than this strategy does.

- *Via Negativa*

Via negativa is a principle that says to continually chip away at sources of downside risk, working to remove the bad instead of increase the good. It also says to avoid games that have obviously large sources of downside risk. The principle here is that downside risk is unavoidable, but by making it a priority to remove sources of downside risks over time, you can significantly improve your chances.

In the bag of marbles example, this might look like getting a magnet that can over time begin to suck all the red marbles/items out of the bag, so you're left with only the positive value marbles. For a more concrete example, this would involve paying off debt before investing in new equipment for a business, even if the rate of return from the new equipment would be higher than the rate of interest on the loan. The loan is a downside risk that could be catastrophic in the case of a black swan that prevented that upside potential from emerging.

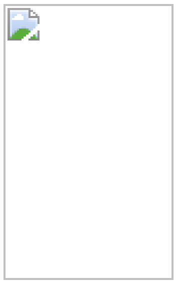
This helps deal with black swans, dynamic environments, and adversarial environments by making sure you don't lose more than you can afford given that the distribution takes a turn for the worse.

- *Skin in the Game*

Skin in the game is a principle that comes from applying anti-fragility on a systems level. It says that in order to encourage individuals and organizations to create anti-fragile systems, they must be exposed to the downside risk that they create.

If I can create downside risk for others that I am not exposed to, I can create a locally anti-fragile environment that nonetheless increases fragility globally. The Skin in the game principle aims to combat two forces that create molochian anti-fragile environments- moral hazards and negative externalities.

Effectuation



Effectuation is a term coined by Saras Sarasvathy to describe a particular type of proactive strategy she found when studying expert entrepreneurs. Instead of looking to mitigate risks by choosing strategies that were flexible in the presence of large downsides risks (antifragility), these entrepreneurs instead worked to shift the distribution such that there were no downside risks, or shift the rules such that the risks were no longer downsides. There's not a book a can recommend that's great at explaining effectuation, but two OK ones are Effectuation by Saras Sarasvathy and Zero to One by Peter Thiel. This 3-page infographic on effectuation is also decent.

Note that Effectuation and Antifragility explicitly trade off against each other. Antifragility trades away certainty for flexibility while Effectuation does the opposite.

In terms of the "marbles in a bag" metaphor, Effectaution can be seen as pouring a lot of marbles that are really helpful to you into the bag, then reaching in and pulling them out.

- *Pilot-in-Plane Principle*

The pilot-in-plane principle is a general way of thinking that says control is better than both prediction and anti-fragility. The pilot-in-plane principle emphasizes proactively shaping risks and rewards, instead of creating a system that can deal with unknown or shifting risks and rewards. The quote that best summarizes this principle is the Peter Drucker quote "The best way to predict the future is to create it."

This principle also isn't much use with black swans. It deals with dynamic environments by seizing control of the forces that shape those dynamic environments. It deals with adversarial environments by shaping the adversarial landscape.

- *Affordable Loss Principle*

The affordable loss principle simply says that you shouldn't risk more than you're willing to lose on any given bet. It's Effectuation's answer to Via Negativa principle.

The difference is that while Via negativa recommends policies that search for situations with affordable downside, and focus on mitigating unavoidable downside, Affordable loss focuses on using your resources to shape situations in which the loss of all parties is affordable.

It's not enough to just make bets you can afford to lose, you have to figure out how to do this while maximizing upside. Can you get a bunch of people to band together to put in a little, so that everyone can afford to lose what they're putting in, but you have a seat at the table? Can you have someone else shoulder the risk who can afford to lose more> Can you get guarantees or insurance to minimize downside risk while still getting the upside? Many of these principles break the Skin in the Game principle needed for anti-fragility, but work perfectly (without calling up Moloch) when using an effectuative strategy. This is the affordable loss principle.

It helps with black swans by creating buffers that protect catastrophic loss. It helps with dynamic environments by keeping what can you lose constant even as the environment changes. It helps with adversarial environments by making sure you can afford to lose to your adversary.

- *Bird-in-Hand Principle*

The bird-in-hand principle says that you should use your existing knowledge, expertise, connections, and resources to shift the distribution in your favor. It also says that you should only choose to play games where you have enough of these existing resources to shift the distribution. Peter Thiel says to ask the question "What do I believe that others do not?" Saras Sarasvathy says to look at who you are, what you know, and who you know.

This helps with Black Swans by preventing some of them from happening. It helps with dynamic environments by seizing control of the process that is causing the environment to change, making most of the change come from you. It helps with adversarial environments by ensuring that you have an unfair advantage in the game.

- *Lemonade Principle*

The lemonade principle that says when the unexpected happens, you should use that as an opportunity to re-evaluate the game you're playing, and seeing if there's a more lucrative game you should be playing instead. Again, the idea of "make the most of a bad situation" might seem obvious, but through the creative and proactive lens of effectuation, it's taken to the extreme. Instead of saying "What changes can I make to my current approach given this new situation?" the lemonade principle says to ask "Given this new situation, what's the best approach to take?"

This helps with Black Swans by using them as lucrative opportunities for gaining utility. It helps with dynamic environments by constantly finding the best opportunity given the current landscape. It helps with adversarial environments by refusing to play losing games.

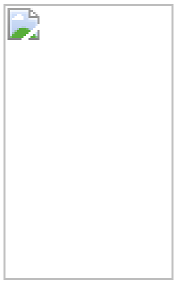
- *Patchwork Quilt Principle*

The patchwork quilt principle says that you should trade flexibility for certainty by bringing on key partners. The partners get to have more of a say in the strategies you use, but in turn you get access to their resources and the certainty that they're on board.

While the original work on effectuation paints this principle as only having to do with partnerships, I like to think of it as a general principle where you should be willing to limit your options if it limits your downside risk and volatility more. The inverse of the optionality principle from antifragile strategies.

This strategy doesn't really help with black swans that much. It helps with dynamic environment by making the environment less dynamic through commitments. It helps with Adversarial environments by turning potential adversaries into allies.

Capability Enhancement



Capability enhancement is a general strategy of trying to improve capabilities such that knightian risks are turned into opaque risks (which are then turned into transparent risks through sampling and modelling). Unlike the previous to ways to mitigate knightian risk, this is more a class of strategies than a strategy in its' own right. In terms of the "marbles in a bag" analogy, capability enhancement might be building x-ray googles to look through the bag, or getting really good at shaking it to figure out the distribution.

Black Swans can be turned opaque by knowing more (and having less unknown unknowns. Dynamic environments can be turned opaque by increasing the speed of sampling or modelling, or the accuracy or applicability of models. Adversarial environments can be turned opaque by giving better strategies to model or face adversaries (and their interactions with each other).

There are numerous classification schemes one could use for all the various types of capability enhancement. Instead of trying to choose one, I'll simply list a few ways that I see people trying to approach this, with no attempt at completeness or consistent levels of abstraction.

- *Personal Psychology Enhancement*

By making people think better, work more, be more effective, an individual can increase the class of problems that become opaque to them. This is one approach that CFAR and Leverage are taking.

- *Better Models*

By creating better models of how the world works, risks that were previously knightian to you become opaque. I would put Leverage, FHI, and MIRI into the class of organizations that are taking this approach to capability enhancement. The sequences could fit here as well.

- *Better Thinking Tools*

By creating tools that can themselves help you model things, you can make risks opaque that were previously Knightian. I would put Verity, Guesstimate, and Roam in this category.

- *Improving Group Dynamics*

By figuring out how to work together better, organizations can turn risks from knightian to opaque. Team Team on Leverage, and CFARs work on group rationality both fit into this category.

- *Collective Intelligence and Crowdsourcing*

By figuring out how to turn a group of people into a single directed agent, you can often shore up individuals weaknesses and amplify their strengths. This allows risks that were previously knightian to individuals become opaque to the collective.

I would put Metaculus, Verity, and LessWrong into this category.

Knightian Risks in Real Life

0 to 1 Companies

When a company is creating something entirely new (in the Wardley Mapping sense), it's taking a Knightian risk. Sampling is fairly useless here because people don't know they want what doesn't exist, and naive approaches to modelling won't work because your inputs are all junk data that exists without your product in the market.

How would each of these strategies handle this situation?

- *Effectuation*

Start your company in an industry where you have pre-existing connections, and in which you have models or information that others don't ("What do you believe that others do not?"). Before building the product, get your contacts to pay up front to get you to build it, therefore limiting risk. If something goes wrong in the building of the product, take all the information you've gathered and the alliances you've already made, and figure out what the best opportunity is with that information and resources.

- *Anti-Fragility*

Create a series of small experiments with prototypes of your products. Keep the ones that succeed, and branch them off into more variations, only keeping the ones that do well. Avoid big contracts like in the effectuation example, only taking small contracts that can let you pivot at a moments notice if needed.

- *Capability Enhancement*

Create a forecasting tournament for the above product variations. Test only the ones that have positive expected value. Over time, you'll have less and less failed experiments as your reputation measures get better. Eventually, you may be able to skip many experiments all together and just trust the forecasting data. *If you're interested in this type of thing we should really chat.*

AGI Risk

At first glance, it seems like many of these strategies such as Effectuation apply more to individual or group risks than global risks. It's not clear for instance how an effectual strategy of shifting the risks to people who can handle them applies on a society wide scale. I do however think that this categorization scheme has something to say about existential risk, and will illustrate with a few examples of ways to mitigate AGI Risk. I recognize that many of these examples are incredibly simplified and unrealistic. The aim is simply to show how this categorization scheme can be used to meaningfully think about existential risk, not to make actual policy suggestions or leaps forward.

How might we mitigate AI risk using the strategies discussed here?

- *Capability Enhancement/Modelling/Sampling*

A capability enhancement/sampling/modelling strategy might be to get a bunch of experts together and forecast how soon we'll get AGI. Then, get a bunch of forecasting experts together and create a function that determines how long it takes to develop benevolent AGI given the amount of AI safety researchers. Finally, create a plan to hire enough AI safety researchers that we develop the ability to create safe AGI before we develop the ability to develop unsafe AGI. If we find that there's simply no way to discover AI safety fast enough given current methods, create tools to get better at working on AI safety. If you find that the confidence intervals on AGI timelines are too wide, create tools that can allow you to narrow them.

- *Anti-fragility*

An anti-fragile strategy might look like developing a system of awareness of AI risk and enough funding such that you can create a strategy where two AI safety researchers are hired for every non-safety AI researcher that is hired. Thus, the more you expose yourself to the existential risk of AGI, the faster you create the mechanism that protects you from that risk. This might be pared with a system that tries different approaches to AI safety, and splits off the groups that are doing the best every few years into two groups, these evolving a system that increases the effectiveness of AI safety researchers over time.

- *Effectuation*

The effectual strategy, instead of taking the timeline for AI as a given, would instead ask "How can we change this timeline such that there's less risk?" Having asked that question, and recognizing that pretty much any answer exists in an adversarial environment, the question becomes "What game can we play that we, as effective altruists, have a comparative advantage at compared to our adversaries?" If the answer is something like "We have an overbundance of smart, capabable people who are willing to forgo both money and power for altruistic reasons," then maybe the game we play is getting a bunch of effective altruists to run for local offices in municipal elections, and influence policy from the ground up by coordinating laws on a municipal level to create a large effect of requiring safety teams for ML teams (among many other small policies). Obviously a ridiculous plan, but it does illustrate how the different risk mitigation strategies can suggest vastly different object level policies.

Exercise for the reader: Robin Hanson worries about a series of catastrophic risks that tax humanity beyond it's resources (I can't find the article to link here but if someone knows it let me know in the comments). We might be able to handle climate change, or an asteroid, or an epidemic on their own, but if by chance they hit together, we pass a critical threshold that we simply can't recover from.

How would you analyze and mitigate this situation of "stacked catastrophic risks" using the framework above?

Thanks to Linda Linsefors for reviewing early drafts.

Subagents, akrasia, and coherence in humans

In my previous posts, I have been building up a model of mind as a collection of subagents with different goals, and no straightforward hierarchy. This then raises the question of how that collection of subagents can exhibit coherent behavior: after all, many ways of aggregating the preferences of a number of agents fail to create consistent preference orderings.

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. If an agent is not coherent in this way, then bad things are likely to happen to them.

Now, we all know that humans sometimes express incoherent behavior. But on the whole, people still do okay: the median person in a developed country still manages to survive until their body starts giving up on them, and typically also manages to have and raise some number of initially-helpless children until *they* are old enough to take care of themselves.

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. The conclusion of this post will be:

> 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.

Correcting your behavior as a default

There are many situations in which we exhibit incoherent behavior simply because we’re not aware of it. For instance, suppose that I do my daily chores in a particular order, when doing them in some other order would save more time. If you point this out to me, I’m likely to just say “oh”, and then adopt the better system.

Similarly, several of the experiments which get people to exhibit incoherent behavior rely on showing different groups of people different formulations of the same question, and then indicating that different framings of the same question get different answers from people. It doesn’t work quite as well if you show the different formulations to the *same* people, because then many of them will realize that differing answers would be inconsistent.

But there are also situations in which someone realizes that they are behaving in a nonsensical way, yet will continue behaving in that way. Since people usually *can* change suboptimal behaviors, we need an explanation for why they sometimes *can’t*.

Towers of protectors as a method for coherence

In my post about Internal Family Systems, I discussed a model of mind composed of several different kinds of subagents. One of them, the default planning subagent, is a module just trying to straightforwardly find the best thing to do and then execute that. On the other hand, *protector* subagents exist to prevent the system from getting into situations which were catastrophic before. If they think that the default planning subagent is doing something which seems dangerous, they will override it and do something else instead. (Previous versions of the IFS post called the default planning agent, “a reinforcement learning subagent”, but this was potentially misleading since several other subagents were reinforcement learning ones too, so I’ve changed the name.)

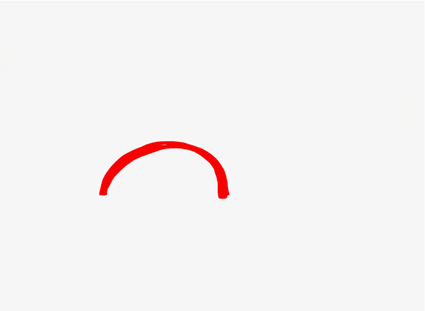
Thus, your behavior can still be coherent even if you *feel* that you are failing to act in a coherent way. You simply don’t realize that a protector is carrying out a routine intended to avoid dangerous outcomes - and this might actually be a very successful way of keeping you out of danger. Some subagents in your mind think that doing X would be a superior strategy, but the protector thinks that it would be a horrible idea - so from the point of view of the system as a whole, doing X is *not* a better strategy, so not switching to it is actually better.

On the other hand, it may also be the case that the protector’s behavior, while keeping you out of situations which the protector considers unacceptable, is causing other outcomes which are *also* unacceptable. The default planning subagent may realize this - but as already established, any protector can overrule it, so this doesn’t help.

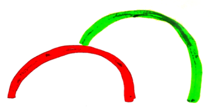
Evolution’s answer here seems to be spaghetti towers. The default planning subagent might *eventually* figure out the better strategy, which avoids both the thing that the protector is trying to block *and* the new bad outcome. But it could be dangerous to wait that long, especially since the default planning agent doesn't have direct access to the protector's goals. So for the same reasons why a separate protector subagent was created to avoid the *first* catastrophe, the mind will create or recruit a protector to avoid the *second* catastrophe - the one that the first protector keeps causing.

With permission, I’ll borrow the illustrations from eukaryote’s spaghetti tower post to illustrate this.

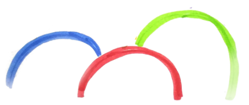
Example Eric grows up in an environment where he learns that disagreeing with other people is unsafe, and that he should always agree to do things that other people ask of him. So Eric develops a protector subagent running a pleasing, submissive behavior.



Unfortunately, while this tactic worked in Eric's childhood home, once he became an adult he starts saying "yes" to too many things, without leaving any time for his own needs. But saying "no" to anything still feels unsafe, so he can't just stop saying "yes". Instead he develops a protector which tries to keep him out of situations where people would ask him to do anything. This way, he doesn't need to say "no", and also won't get overwhelmed by all the things that he has promised to do. The two protectors together form a composite strategy.



While this helps, it still doesn't entirely solve the issue. After all, there are plenty of reasons that might push Eric into situations where someone would ask something of him. He still ends up agreeing to do lots of things, to the point of neglecting his own needs. Eventually, his brain creates another protector subagent. This one causes exhaustion and depression, so that he now has a socially-acceptable reason for being unable to do all the things that he has promised to do. He continues saying "yes" to things, but also keeps apologizing for being unable to do things that he (honestly) intended to do as promised, and eventually people realize that you probably shouldn't ask him to do anything that's really important to get done.



And while this kind of a process of stacking protector on top of a protector is not perfect, for most people it mostly works out okay. Almost everyone ends up having their unique set of minor neuroses and situations where they don't quite behave rationally, but as they learn to understand themselves better, their default planning subagent gets better at working around those issues. This might also make the various protectors relax a bit, since the various threats are generally avoided and there isn't a need to keep avoiding them.

Gradually, as negative consequences to different behaviors become apparent, behavior gets adjusted - either by the default planning subagents or by spawning more protectors - and remains coherent overall.

But sometimes, especially for people in highly stressful environments where almost any mistake may get them punished, or when they end up in an environment that their old tower of protectors is no longer well-suited for (distributional shift), things don't go as well. In that situation, their minds may end up looking like this a hopelessly tangled web, where they have almost no flexibility. Something happens in their environment, which sets off one protector, which sets off another, which sets off another - leaving them with no room for flexibility or rational planning, but rather forcing them to act in a way which is almost bound to only make matters worse.

This kind of an outcome is obviously bad. So besides building spaghetti towers, the second strategy which the mind has evolved to employ for keeping its behavior coherent while piling up protectors, is the ability to re-process memories of past painful events.

As I discussed in my original IFS post, the mind has methods for bringing up the original memories which caused a protector to emerge, in order to re-analyze them. If ending up in some situation is actually no longer catastrophic (for instance, you are no longer in your childhood home where you get punished simply for not wanting to do something), then the protectors which were focused on avoiding that outcome can relax and take a less extreme role.

For this purpose, there seems to be a built-in tension. Exiles (the IFS term for subagents containing memories of past trauma) "want" to be healed and will do things like occasionally sending painful memories or feelings into consciousness so as to become the center of attention, especially if there is something about the current situation which resembles the past trauma. This also acts as what my IFS post called a fear model - something that warns of situations which resemble the past trauma enough to be considered dangerous in their own right. At the same time, protectors "want" to keep the exiles hidden and inactive, doing anything that they can for keeping them so. Various schools of therapy - IFS one of them - seek to tap into this existing tension so as to reveal the trauma, trace it back to its original source, and heal it.

Coherence and conditioned responses

Besides the presence of protectors, another possibility for why we might fail to change our behavior are strongly conditioned habits. Most human behavior involves automatic habits: behavioral routines which are triggered by some sort of a cue in the environment, and lead to or have once led to a reward. (Previous discussion; see also.)

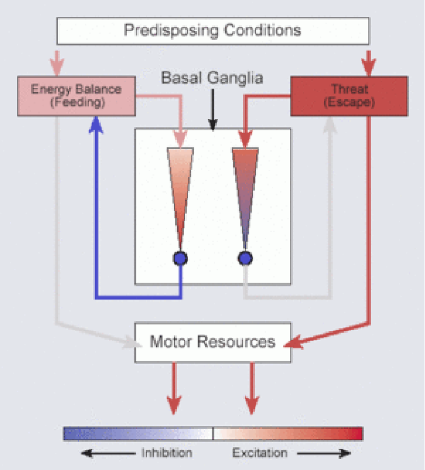
The problem with this is that people might end up with habits that they wouldn’t want to have. For instance, I might develop a habit of checking social media on their phone when I’m bored, creating a loop of boredom (cue) -> looking at social media (behavior) -> seeing something interesting on social media (reward).

Reflecting on this behavior, I notice that back when I *didn’t* do it, my mind was more free to wander when I was bored, generating motivation and ideas. I think that my old behavior was more valuable than my new one. But even so, my new behavior still delivers enough momentary satisfaction to keep reinforcing the habit.

Subjectively, this feels like an increasing compulsion to check my phone, which I try to resist since I know that long-term it would be a better idea to not be checking my phone all the time. But as the compulsion keeps growing stronger and stronger, eventually I give up and look at the phone anyway.

The exact neuroscience of what is happening at such a moment remains only partially understood (Simpson & Balsam 2016). However, we know that whenever different subsystems in the brain produce conflicting motor commands, that conflict needs to be resolved, with only one at a time being granted access to the “final common motor path”. This is thought to happen in the basal ganglia, a part of the brain closely involved in action selection and connected to the global neuronal workspace.

One model (e.g. Redgrave 2007, McHaffie 2005) is that the basal ganglia receives inputs from many different brain systems; each of those systems can send different “bids” supporting or opposing a specific course of action to the basal ganglia. A bid submitted by one subsystem may, through looped connections going back from the basal ganglia, inhibit other subsystems, until one of the proposed actions becomes sufficiently dominant to be taken.



The above image from Redgrave 2007 has a conceptual image of the model, with two example subsystems shown. Suppose that you are eating at a restaurant in Jurassic Park when two velociraptors charge in through the window. Previously, your hunger system was submitting successful bids for the “let’s keep eating” action, which then caused inhibitory impulses to the be sent to the threat system. This inhibition prevented the threat system from making bids for silly things like jumping up from the table and running away in a panic. However, as your brain registers the new situation, the threat system gets significantly more strongly activated, sending a strong bid for the “let’s run away” action. As a result of the basal ganglia receiving that bid, an inhibitory impulse is routed from the basal ganglia to the subsystem which was previously submitting bids for the “let’s keep eating” actions. This makes the threat system’s bids even stronger relative to the (inhibited) eating system’s bids.

Soon the basal ganglia, which was previously inhibiting the threat subsystem’s access to the motor system while allowing the eating system access, withdraws that inhibition and starts inhibiting the eating system’s access instead. The result is that you jump up from your chair and begin to run away. Unfortunately, this is hopeless since the velociraptor is faster than you. A few moments later, the velociraptor’s basal ganglia gives the raptor’s “eating” subsystem access to the raptor’s motor system, letting it happily munch down its latest meal.

But let’s leave velociraptors behind and go back to our original example with the phone. Suppose that you have been trying to replace the habit of looking at your phone when bored, to instead smiling and directing your attention to pleasant sensations in your body, and then letting your mind wander.

Until the new habit establishes itself, the two habits will compete for control. Frequently, the old habit will be stronger, and you will just automatically check your phone without even remembering that you were supposed to do something different. For this reason, behavioral change programs may first spend several weeks just practicing *noticing* the situations in which you engage in the old habit. When you *do* notice what you are about to do, then more goal-directed subsystems may send bids towards the “smile and look for nice sensations” action. If this happens and you pay attention to your experience, you may notice that long-term it actually feels more pleasant than looking at the phone, reinforcing the new habit until it becomes prevalent.

To put this in terms of the subagent model, we might drastically simplify things by saying that the neural pattern corresponding to the old habit is a subagent reacting to a specific sensation (boredom) in the consciousness workspace: its reaction is to generate an intention to look at the phone. At first, you might train the subagent responsible for monitoring the contents of your consciousness, to output moments of introspective awareness highlighting when that intention appears. That introspective awareness helps alert a goal-directed subagent to try to trigger the new habit instead. Gradually, a neural circuit corresponding to the new habit gets trained up, which starts sending its own bids when it detects boredom. Over time, reinforcement learning in the basal ganglia starts giving that subagent’s bids more weight relative to the old habit’s, until it no longer needs the goal-directed subagent’s support in order to win.

Now this model helps incorporate things like the role of having a vivid emotional motivation, a sense of hope, or psyching yourself up when trying to achieve habit change. Doing things like imagining an outcome that you wish the habit to lead to, may activate additional subsystems which care about those kinds of outcomes, causing them to submit additional bids in favor of the new habit. The extent to which you succeed at doing so, depends on the extent to which your mind-system considers it *plausible* that the new habit leads to the new outcome. For instance, if you imagine your exercise habit making you strong and healthy, then subagents which care about strength and health might activate to the extent that you believe this to be a likely outcome, sending bids in favor of the exercise action.

On this view, one way for the mind to maintain coherence and readjust its behaviors, is its ability to re-evaluate old habits in light of which subsystems get activated when reflecting on the possible consequences of new habits. An old habit having been strongly reinforced reflects that a great deal of evidence has accumulated in favor of it being beneficial, but the behavior in question can still be overridden if enough influential subsystems weigh in with their evaluation that a new behavior would be more beneficial in expectation.

Some subsystems having concerns (e.g. immediate survival) which are ranked more highly than others (e.g. creative exploration) means that the decision-making process ends up carrying out an implicit expected utility calculation. The strengths of bids submitted by different systems do not just reflect the probability that those subsystems put on an action being the most beneficial. There are also different mechanisms giving the bids from different subsystems varying amounts of weight, depending on how important the concerns represented by that subsystem happen to be in that situation. This ends up doing something like weighting the probabilities by utility, with the kinds of utility calculations that are chosen by evolution and culture in a way to maximize genetic fitness on average. Protectors, of course, are subsystems whose bids are weighted particularly strongly, since the system puts high utility on avoiding the kinds of outcomes they are trying to avoid.

The original question which motivated this section was: why are we sometimes incapable of adopting a new habit or abandoning an old one, despite knowing that to be a good idea? And the answer is: because we *don’t* know that such a change would be a good idea. Rather, *some* subsystems *think* that it would be a good idea, but other subsystems remain unconvinced. Thus the system’s overall judgment is that the old behavior should be maintained.

Interlude: Minsky on mutually bidding subagents

> I was trying to concentrate on a certain problem but was getting bored and sleepy. Then I imagined that one of my competitors, Professor Challenger, was about to solve the same problem. An angry wish to frustrate Challenger then kept me working on the problem for a while. The strange thing was, this problem was not of the sort that ever interested Challenger.
>
> What makes us use such roundabout techniques to influence ourselves? Why be so indirect, inventing misrepresentations, fantasies, and outright lies? Why can't we simply tell ourselves to do the things we want to do? [...]
>
> Apparently, what happened was that my agency for Work exploited Anger to stop Sleep. But why should Work use such a devious trick?
>
> To see why we have to be so indirect, consider some alternatives. If Work could simply turn off Sleep, we'd quickly wear our bodies out. If Work could simply switch Anger on, we'd be fighting all the time. Directness is too dangerous. We'd die.

> Extinction would be swift for a species that could simply switch off hunger or pain. Instead, there must be checks and balances. We'd never get through one full day if any agency could seize and hold control over all the rest. This must be why our agencies, in order to exploit each other's skills, have to discover such roundabout pathways. All direct connections must have been removed in the course of our evolution.
>
> This must be one reason why we use fantasies: to provide the missing paths. You may not be able to make yourself angry simply by deciding to be angry, but you can still imagine objects or situations that make you angry. In the scenario about Professor Challenger, my agency Work exploited a particular memory to arouse my Anger's tendency to counter Sleep. This is typical of the tricks we use for self-control.
>
> Most of our self-control methods proceed unconsciously, but we sometimes resort to conscious schemes in which we offer rewards to ourselves: "If I can get this project done, I'll have more time for other things." However, it is not such a simple thing to be able to bribe yourself. To do it successfully, you have to discover which mental incentives will actually work on yourself. This means that you - or rather, your agencies - have to learn something about one another's dispositions. In this respect the schemes we use to influence ourselves don't seem to differ much from those we use to exploit other people - and, similarly, they often fail. When we try to induce ourselves to work by offering ourselves rewards, we don't always keep our bargains; we then proceed to raise the price or even deceive ourselves, much as one person may try to conceal an unattractive bargain from another person.
>
> Human self-control is no simple skill, but an ever-growing world of expertise that reaches into everything we do. Why is it that, in the end, so few of our self-incentive tricks work well? Because, as we have seen, directness is too dangerous. If self-control were easy to obtain, we'd end up accomplishing nothing at all.

-- Marvin Minsky, *The Society of Mind*

Akrasia is subagent disagreement

You might feel that the above discussion doesn't still entirely resolve the original question. After all, sometimes we *do* manage to change even strongly conditioned habits pretty quickly. Why is it sometimes hard and sometimes easier?

Redgrave et al. (2010) discuss two modes of behavioral control: goal-directed versus habitual. Goal-directed control is a relatively slow mode of decision-making, where "action selection is determined primarily by the relative utility of predicted outcomes", whereas habitual control involves more directly conditioned stimulus-response behavior. Which kind of subsystem is in control is complicated, and depends on a variety of factors (the following quote has been edited to remove footnotes to references; see the original for those):

> Experimentally, several factors have been shown to determine whether the agent (animal or human) operates in goal-directed or habitual mode. The first is over-training: here, initial control is largely goal-directed, but with consistent and repeated training there is a gradual shift to stimulus–response, habitual control. Once habits are established, habitual responding tends to dominate, especially in stressful situations in which quick reactions are required. The second related

> factor is task predictability: in the example of driving, talking on a mobile phone is fine so long as everything proceeds predictably. However, if something unexpected occurs, such as someone stepping out into the road, there is an immediate switch from habitual to goal-directed control. Making this switch takes time and this is one of the reasons why several countries have banned the use of mobile phones while driving. The third factor is the type of reinforcement schedule: here, fixed-ratio schedules promote goal-directed control as the outcome is contingent on responding (for example, a food pellet is delivered after every n responses). By contrast, interval schedules (for example, schedules in which the first response following a specified period is rewarded) facilitate habitual responding because contingencies between action and outcome are variable. Finally, stress, often in the form of urgency, has a powerful influence over which mode of control is used. The fast, low computational requirements of stimulus–response processing ensure that habitual control predominates when circumstances demand rapid reactions (for example, pulling the wrong way in an emergency when driving on the opposite side of the road). Chronic stress also favours stimulus–response, habitual control. For example, rats exposed to chronic stress become, in terms of their behavioural responses, insensitive to changes in outcome value and resistant to changes in action–outcome contingency. [...]
>
> Although these factors can be seen as promoting one form of instrumental control over the other, real-world tasks often have multiple components that must be performed simultaneously or in rapid sequences. Taking again the example of driving, a driver is required to continue steering while changing gear or braking. During the first few driving lessons, when steering is not yet under automatic stimulus–response control, things can go horribly awry when the new driver attempts to change gears. By contrast, an experienced (that is, ‘over-trained’) driver can steer, brake and change gear automatically, while holding a conversation, with only fleeting contributions from the goal-directed control system. This suggests that many skills can be deconstructed into sequenced combinations of both goal-directed and habitual control working in concert. [...]
>
> Nevertheless, a fundamental problem remains: at any point in time, which mode should be allowed to control which component of a task? Daw et al. have used a computational approach to address this problem. Their analysis was based on the recognition that goal-directed responding is flexible but slow and carries comparatively high computational costs as opposed to the fast but inflexible habitual mode. They proposed a model in which the relative uncertainty of predictions made by each control system is tracked. In any situation, the control system with the most accurate predictions comes to direct behavioural output.

Note those last sentences: besides the subsystems making their own predictions, there might also be a meta-learning system keeping track of which other subsystems tend to make the most accurate predictions in each situation, giving extra weight to the bids of the subsystem which has tended to perform the best in that situation. We’ll come back to that in future posts.

This seems compatible with my experience in that, I feel like it’s possible for me to change even entrenched habits relatively quickly - *assuming that the new habit really is unambiguously better*. In that case, while I might forget and lapse to the old habit a few times, there’s still a rapid feedback loop which quickly indicates that the goal-directed system is simply *right* about the new habit being better.

Or, the behavior in question might be sufficiently complex and I might be sufficiently inexperienced at it, that the goal-directed (default planning) subagent has always mostly remained in control of it. In that case change is again easy, since there is no strong habitual pattern to override.

In contrast, in cases where it's hard to establish a new behavior, there tends to be some kind of genuine uncertainty:

- The benefits of the old behavior have been validated in the form of direct experience (e.g. unhealthy food that tastes good, has in fact tasted good each time), whereas the benefits of the new behavior come from a less trusted information source which is harder to validate (e.g. I've read scientific studies about the long-term health risks of this food).
- Immediate vs. long-term rewards: the more remote the rewards, the larger the risk that they will for some reason never materialize.
- High vs. low variance: sometimes when I'm bored, looking at my phone produces genuinely *better* results than letting my thoughts wander. E.g. I might see an interesting article or discussion, which gives me novel ideas or insights that I would not otherwise have had. Basically looking at my phone usually produces worse results than not looking at it - but sometimes it also produces much better ones than the alternative.
- Situational variables affecting the value of the behaviors: looking at my phone can be a way to escape uncomfortable thoughts or sensations, for which purpose it's often excellent. This then also tends to reinforce the behavior of looking at the phone when I'm in the same situation otherwise, but *without* uncomfortable sensations that I'd like to escape.

When there is significant uncertainty, the brain seems to fall back to those responses which have worked the best in the past - which seems like a reasonable approach, given that intelligence involves hitting tiny targets in a huge search space, so most novel responses are likely to be wrong.

As the above excerpt noted, the tendency to fall back to old habits is exacerbated during times of stress. The authors attribute it to the need to act quickly in stressful situations, which seems correct - but I would also emphasize the fact that negative emotions in general tend to be signs of something being wrong. E.g. Eldar et al. (2016) note that positive or negative moods tend to be related to whether things are going better or worse than expected, and suggest that mood is a *computational representation of momentum*, acting as a sort of global update to our reward expectations.

For instance, if an animal finds more fruit than it had been expecting, that may indicate that spring is coming. A shift to a good mood and being "irrationally optimistic" about finding fruit even in places where the animal hasn't seen fruit in a while, may actually serve as a rational pre-emptive update to its expectations. In a similar way, things going less well than expected may be a sign of some more general problem, necessitating fewer exploratory behaviors and less risk-taking, so falling back into behaviors for which there is a higher certainty of them working out.

So to repeat the summary that I had in the beginning: 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 protector whose bids get a lot of weight) 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.

mAIry's room: AI reasoning to solve philosophical problems

This post grew out of a conversation with Laurent Orseau; we were initially going to write a paper for a consciousness/philosophy journal of some sort, but that now seems unlikely, so I thought I'd post the key ideas here.

A summary of this post can be found here - it even has some diagrams.

The central idea is that thinking in terms of AI or similar artificial agent, we can get some interesting solutions to old philosophical problems, such as the Mary's room/knowledge problem. In essence, simple agents exhibit similar features to Mary in the thought experiments, so (most) explanations of Mary's experience must also apply to simple artificial agents.

To summarise:

- Artificial agents can treat certain inputs as if the input were different from mere information.
- This analogises loosely to how humans "experience" certain things.
- If the agent is a more limited (and more realistic) design, this analogy can get closer.
- There is an artificial version of Mary, mAIry, which would plausibly have something similar to what Mary experiences within the thought experiment.

Edit: See also orthonormal's sequence here.

Mary's Room and the Knowledge problem

In this thought experiment, Mary has been confined to a grey room from birth, exploring the outside world only through a black-and-white monitor.

Though isolated, Mary is a brilliant scientist, and has learnt all there is to know about light, the eye, colour theory, human perception, and human psychology. It would seem that she has all possible knowledge that there could be about colour, despite having never seen it.

Then one day she gets out of her room, and says "wow, so that's what purple looks like!".

Has she learnt anything new here? If not, what is her exclamation about? If so, what is this knowledge - Mary was supposed to know everything there was to know about colour already?

Incidentally, I chose "purple" as the colour Mary would see, as the two colours most often used, red and blue, lead to the confusion as to what "seeing red/blue" means - is this about the brain, or is it about the cones in the eye. But seeing purple is strictly about perception in the brain.

Example in practice

Interestingly, there are real example of Mary's room-like situations. Some people with red-green colour-blindness can suddenly start seeing new colours with the right glasses. Apparently this happens because the red and green cones in their eyes are almost identical, so tend to always fire together. But "almost" is not "exactly", and the glasses force green and red colours apart, so the red and green cones start firing separately, allowing the colour blind to see or distinguish new colours.

Can you feel my pain? The AI's reward channel

This argument was initially presented here.

AIXI

Let's start with the least human AI we can imagine: AIXI, which is more an equation than an agent. Because we'll be imagining multiple agents, let's pick any computable version of AIXI, such as AIXItl.

There will be two such AIXItl's, called A_r and A_q, and they will share observations and rewards: at turn i, this will be o_i, r_i, and q_i, with r_i the reward of A_r and q_i the reward of A_q.

To simplify, we'll ignore the game theory between the agents; each agent will treat the other as part of the environment and attempt to maximise their reward around this constraint.

Then it's clear that, even though r_i and q_i are both part of each agent's observation, each agent will treat their own reward in a special way. Their actions are geared to increasing their own reward; A_r might find q_i informative, but has no use for it beyond that.

For example, A_r might sacrifice current r_i to get information that could lead it to increase $r_{j>i}$; it would never do so to increase $q_{j>i}$. It would sacrifice all q-rewards to increase the expected sum of r_i; indeed it would sacrifice its knowledge of q_i entirely to increase that expected sum by the tiniest amount. And A_q would be in the exact opposite situation.

The A_r agent would also do other things, like sacrificing r_i in counterfactual universes to increase r_i in this one. It would also refuse the following trade: perfect knowledge of the ideal policy that would have maximised expected r_i, in exchange for the r_i being set to 0 from then on. In other words, it won't trade r_i for perfect information about r_i.

So what are these reward channels to these agents? It would go too far to call them qualia, but they do seem to have some features of pleasure/pain in humans. We don't feel the pleasure and pain of others in the same way we feel our own. We don't feel counterfactual pain as we feel real pain; and we certainly wouldn't agree to suffer maximal pain in exchange for knowing how we could have otherwise felt maximal pleasure. Pleasure and pain can motivate us to action in ways that few other things can: we don't treat them as pure information.

Similarly, the A_r doesn't treat r_i purely as information either. To stretch the definition of a word, we might say that A_r is *experiencing* r_i in ways that it doesn't experience q_i or o_i.

Let's try and move towards a more human-like agent.

TD-Lambda learning

TD stands for temporal difference learning: learning by the difference between a predicted reward and the actual reward. For the TD-Lambda algorithm, the agent uses $V(s)$: the estimated value of the state s. It then goes on its merry way, and as it observes histories of the form $\ldots s_{i-1} a_{i-1} r_{i-1} s_i a_i r_i s_{i+1} a_{i+1} r_{i+1}$, it updates is estimate of all its past $V(s_i)$ (with a discount factor of $0 \leq \lambda \leq 1$ for more distant past states $s_{j<i}$).

Again, imagine there are two agents, T_r and T_q, with separate reward functions r and q, and that each agent gets to see the other's reward.

What happens when T_r encounters an unexpectedly large or small value of q_i? Well, how would it interpret the q_i in the first place? Maybe as part of the state-data s_{i+1}. In that case, an unexpected q_i moves T_r to a new, potentially unusual state s_{i+1}, rather than an expected s_{i+1}'. But this is only relevant if $V(s_{i+1})$ is very different from $V(s_{i+1}')$: in other words, unexpected q_i are only relevant if they imply something about expected r_i. And even when they do, their immediate impact is rather small: a different state reached.

Compare what happens when T_r encounters an unexpectedly large or small value of r_i . The impact of that is immediate: the information percolates backwards, updating all the $V(s_{j<i})$. There is an immediate change to the inner variables all across the agent's brain.

In this case, the 'experience' of the T_r agent encountering high/low r_i resembles our own experience of extreme pleasure/pain: immediate involuntary re-wiring and change of estimates through a significant part of our brain.

We could even give T_r a certain way of 'knowing' that high/low r_i might be incoming; maybe there's a reliability score for $V(s_i)$, or some way of tracking variance in the estimate. Then a low reliability or high variance score could indicate to the T_r that high/low r_i might happen (maybe these could feed into the learning rate α). But, even if the magnitude of the r_i is not unexpected, it will still cause changes across all the previous estimates - even if these changes are in some sense expected.

mAIry in its room

So we've established that artificial agents can treat certain classes of inputs in a special way, "experiencing" their data (for lack of a better word) in a way that is different from simple information. And sometimes these inputs can strongly rewire the agent's brain/variable values.

Let's now turn back to the initial thought experiment, and posit that we have a mAIry, an AI version of Mary, similarly brought up without the colour purple. mAIry stores knowledge as weights in a neural net, rather than connections of neurons, but otherwise the thought experiment is very similar.

mAIry knows everything about light, cameras, and how neural nets interpret concepts, including colour. It knows that, for example, "seeing purple" corresponds to a certain pattern of activation in the neural net. We'll simplify, and just say that there's a certain node n_p such that, if its activation reaches a certain threshold, the net has "seen purple". mAIry is aware of this fact, and can identify the n_p node within itself, and perfectly predict the sequence of stimuli that could activate it.

If mAIry is still a learning agent, then seeing a new stimuli for the first time is likely to cause a lot of changes in the weights in its nodes; again, these are changes that mAIry can estimate and predict. Let c_p be a Boolean corresponding to whether these changes have happened or not.

What dreams of purple may come...

A sufficiently smart mAIry might be able to force itself to "experience" seeing purple, without ever having seen it. If it has full self-modification powers, it could manually activate n_p and cause the changes that result in c_p being true. With more minor abilities, it could trigger some low-level neurons that caused a similar change in its neural net.

In terms of the human Mary, these would correspond to something like self-brain surgery and self-hypnosis (or maybe self-induced dreams of purple).

Coming out of the room: the conclusion

So now assume that mAIry exits the room for the first time, and sees something purple. It's possible that mAIry has successfully self-modified to activate n_p and set c_p to true. In that case, upon seeing something purple, mAIry gets no extra information, no extra knowledge, and nothing happens in its brain that could correspond to a "wow".

But what if mAIry has not been able to self-modify? Then upon seeing a purple flower, the node n_p is strongly activated for the first time, and a whole series of weight changes flow across mAIry's brain, making c_p true.

That is the "wow" moment for mAIry. Both mAIry and Mary have experienced something; something they both perfectly predicted ahead of time, but something that neither could trigger ahead of time, nor prevent from happening when they did see something purple. The novel activation of n_p and the changes labelled by c_p were both predictable and unavoidable for a smart mAIry without self-modification abilities.

At this point the analogy I'm trying to draw should be clear: activating n_p and the unavoidable changes in the weights that causes c_p to be true, are similar to what a TD-Lambda agent goes through when encountering unexpectedly high or low rewards. They are a "mental experience", unprecedented for the agent even if entirely predictable.

But they are not evidence for epiphenomenalism or against physicalism - unless we want to posit that mAIry is non-physical or epiphenomenal.

It is interesting, though, that this argument suggests that qualia are very real, and distinct from pure information, though still entirely physical.

The Main Sources of AI Risk?

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

There are so many causes or sources of AI risk that it's getting hard to keep them all in mind. I propose we keep a list of the main sources (that we know about), such that we can say that if none of these things happen, then we've mostly eliminated AI risk (as an existential risk) at least as far as we can determine. Here's a list that I spent a couple of hours enumerating and writing down. Did I miss anything important?

1. Insufficient time/resources for AI safety (for example caused by intelligence explosion or AI race)
2. Insufficient global coordination, leading to the above
3. Misspecified or incorrectly learned goals/values
4. Inner optimizers
5. ML differentially accelerating easy to measure goals
6. Paul Christiano's "influence-seeking behavior" (a combination of 3 and 4 above?)
7. AI generally accelerating intellectual progress in a wrong direction (e.g., accelerating unsafe/risky technologies more than knowledge/wisdom about how to safely use those technologies)
8. Metaethical error
9. Metaphilosophical error
10. Other kinds of philosophical errors in AI design (e.g., giving AI a wrong prior or decision theory)
11. Other design/coding errors (e.g., accidentally putting a minus sign in front of utility function, supposedly corrigible AI not actually being corrigible)
12. Doing acausal reasoning in a wrong way (e.g., failing to make good acausal trades, being acausally extorted, failing to acausally influence others who can be so influenced)
13. Human-controlled AIs ending up with wrong values due to insufficient "metaphilosophical paternalism"
14. Human-controlled AIs causing ethical disasters (e.g., large scale suffering that can't be "balanced out" later) prior to reaching moral/philosophical maturity
15. Intentional corruption of human values
16. Unintentional corruption of human values
17. Mind crime (disvalue unintentionally incurred through morally relevant simulations in AIs' minds)
18. Premature value lock-in (i.e., freezing one's current conception of what's good into a utility function)
19. Extortion between AIs leading to vast disvalue
20. Distributional shifts causing apparently safe/aligned AIs to stop being safe/aligned
21. Value drift and other kinds of error as AIs self-modify, or AIs failing to solve value alignment for more advanced AIs
22. Treacherous turn / loss of property rights due to insufficient competitiveness of humans & human-aligned AIs
23. Gradual loss of influence due to insufficient competitiveness of humans & human-aligned AIs
24. Utility maximizers / goal-directed AIs having an economic and/or military competitive advantage due to relative ease of cooperation/coordination, defense

against value corruption and other forms of manipulation and attack, leading to one or more of the above

25. In general, the most competitive type of AI being too hard to align or to safely use
26. Computational resources being too cheap, leading to one or more of the above

(With this post I mean to (among other things) re-emphasize the disjunctive nature of AI risk, but this list isn't fully disjunctive (i.e., some of the items are subcategories or causes of others), and I mostly gave a source of AI risk its own number in the list if it seemed important to make that source more salient. Maybe once we have a list of everything that is important, it would make sense to create a graph out of it.)

Added on 6/13/19:

27. Failure to learn how to deal with alignment in the many-humans, many-AIs case even if single-human, single-AI alignment is solved (suggested by William Saunders)
28. Economics of AGI causing concentration of power amongst human overseers
29. Inability to specify any ‘real-world’ goal for an artificial agent (suggested by Michael Cohen)
30. AI systems end up controlled by a group of humans representing a small range of human values (ie. an ideological or religious group that imposes values on everyone else) (suggested by William Saunders)

Added on 2/3/2020:

31. Failing to solve the commitment races problem, i.e. building AI in such a way that some sort of disastrous outcome occurs due to unwise premature commitments (or unwise hesitation in making commitments!). This overlaps significantly with #27, #19, and #12.

Added on 3/11/2020:

32. Demons in imperfect search (similar, but distinct from, inner optimizers.) See here for illustration.

Added on 10/4/2020:

33. Persuasion tools or some other form of narrow AI leads to a massive deterioration of collective epistemology, dooming humanity to stumble inexorably into some disastrous end or other.

Added on 8/31/2021:

34. Vulnerable world type 1: narrow AI enables many people to destroy world, e.g. R&D tools that dramatically lower the cost for building WMD's.
35. Vulnerable world 2a: We end up with many powerful actors able and incentivized to create civilization-devastating harms.

[Edit on 1/28/2020: This list was created by Wei Dai. Daniel Kokotajlo offered to keep it updated and prettify it over time, and so was added as a coauthor.]

Understanding information cascades

Meta: Because we think understanding info cascades are important, we recently spent ~10 hours trying to figure out how to quantitatively model them, and have contributed our thinking as answers below. While we currently didn't have the time to continue exploring, we wanted to experiment with seeing how much the LW community could together build on top of our preliminary search, so we've put up a basic prize for more work and tried to structure the work around a couple of open questions. This is an experiment! We're looking forward to reading any of your contributions to the topic, including things like summaries of existing literature and building out new models of the domain.

Background

Consider the following situation:

> Bob is wondering whether a certain protein injures the skeletal muscle of patients with a rare disease. He finds a handful papers with some evidence for the claim (and some with evidence against it), so he simply states the claim in his paper, with some caution, and adds that as a citation. Later, Alice comes across Bob's paper and sees the cited claim, and she proceeds to cite Bob, but without tracing the citation trail back to the original evidence. This keeps happening, in various shapes and forms, and after a while a literature of hundreds of papers builds up where it's common knowledge that β amyloid injures the skeletal muscle of patients with inclusion body myositis -- without the claim having accumulated any more evidence. (This real example was taken from Greenberg, 2009, which is a case study of this event.)

An information-cascade occurs when people update on each others beliefs, rather than sharing the causes of those beliefs, and those beliefs end up with a vestige of support that far outstrips the evidence for them. Satvik Beri might describe this as the problem of only sharing the outputs of your thinking process, not your inputs.

The dynamics here are perhaps reminiscent of those underlying various failures of collective rationality such as asset bubbles, bystander effects and stampedes.

Note that his effect is different from other problems of collective rationality like the replication crisis, which involve *low standards* for evidence (such as unreasonably lax p-value thresholds or coordination problems preventing publishing of failed experiments), or the degeneracy of much online discussion, which involves tribal signalling and UI encouraging problematic selection effects. Rather, information cascades involve people *rationally updating* without *any* object-level evidence at all, and would persist even if the replication crisis and online outrage culture disappeared. If nobody lies or tells untruths, you can still be subject to an information cascade.

Questions

Ben and I are confused about how to think about the negative effects of this problem. We understand the basic idea, but aren't sure how to reason quantitatively about the impacts, and how to trade-off solving these problems in a community versus doing other improvements to overall efficacy and efficiency of a community. We currently know only how to think about these qualitatively.

We're posting a couple of related questions that we have some initial thoughts on, that might help clarify the problem.

- How common, and how large, are info-cascades in communities that seek to make intellectual progress, such as academia?
- How can we quantify the impact (harm) of info-cascades, and think about them in cost-effectiveness terms?
- What have been some historically effective ways of responding to cascades, and where have those approaches failed?
- How do you mathematically formalise information cascades around continuous variables?

If you have something you'd like to contribute, but that doesn't seem to fit into the related questions above, leave it as an answer to this question.

Bounties

We are committing to pay at least **either $800 or (No. of answers and comments * $25), whichever is smaller,** for work on this problem recorded on LW, done before May 13th. The prize pool will be split across comments in accordance with how valuable we find them, and we might make awards earlier than the deadline (though if you know you'll put in work in x weeks, it would be good to mention that to one of us via PM).

Ben and Jacob are each responsible for half of the prize money.

Jacob is funding this through Metaculus AI, a new forecasting platform tracking and improving the state-of-the-art in AI forecasting, partly to help avoid info-cascades in the AI safety and policy communities (we're currently live and inviting beta-users, you can sign-up here).

Examples of work each of us are especially excited about:

Jacob

- Contributions to our Guesstimate model (linked here), such as reducing uncertainty on the inputs or using better models.
- Extensions of the Guesstimate model beyond biomedicine, especially in ways that make it more directly applicable to the rationality/effective altruism communities
- Examples and analysis of existing interventions that deal with this and what makes them work, possibly suggestions for novel ones (though avoiding the trap of optimising for good-seeming ideas)
- Discussion of how the problem of info-cascades relates to forecasting

Ben

- Concise summaries of relevant papers and their key contributions
- Clear and concise explanations of what other LWers have found (e.g. turning 5 long answers into 1 medium sized answer that links back to the others while still conveying the key info. Here’s a good example of someone distilling an answer section).

How large is the harm from info-cascades? [Info-cascade series]

This is a question in the info-cascade question series. There is a prize pool of up to $800 for answers to these questions. See the link above for full background on the problem (including a bibliography) as well as examples of responses we'd be especially excited to see.

___

How can we quantify the impact (harm) of info-cascades?

There are many ways in which info-cascades are harmful. Insofar as people base their decisions on the cascaded info, this can result in bad career choices, mistaken research directions, misallocation of grants, a culture that is easier to hijack by cleverly signalling outsiders (by simply "joining the bubble"), and more.

But in order to properly allocate resources to work on info-cascades we need a better model of how large the effects are, and how they compare with other problems. How can we think about info-cascades from a cost-effectiveness perspective?

We are especially interested in answers to this question that ultimately bear on the effective altruism/rationality communities, or analyses of other institutions with insights that transfer to these communities.

As an example step in this direction, we built a Guesstimate model, which is described in an answer below.

Alignment Newsletter #49

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

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

Highlights

Exploring Neural Networks with Activation Atlases *(Shan Carter et al)*: Previous work by this group of people includes The Building Blocks of Interpretability and Feature Visualization, both of which apparently came out before this newsletter started so I don't have a summary to point to. Those were primarily about understanding what individual neurons in an image classifer were responding to, and the key idea was to "name" each neuron with the input that would maximally activate that neuron. This can give you a global view of what the network is doing.

However, such a global view makes it hard to understand the interaction between neurons. To understand these, we can look at a specific input image, and use techniques like attribution. Rather than attribute final classifications to the input, you could attribute classifications to neurons in the network, and then since individual neurons now had meanings (roughly: "fuzzy texture neuron", "tennis ball neuron", etc) you can gain insight to how the network is making decisions *for that specific input*.

However, ideally we would like to see how the network uses interactions between neurons to make decisions in general; not on a single image. This motivates activation atlases, which analyze the activations of a network on a *large dataset* of inputs. In particular, for each of a million images, they randomly choose a non-border patch from the image, and compute the activation vector at a particular layer of the network at that patch. This gives a dataset of a million activation vectors. They use standard dimensionality reduction techniques to map each activation vector into an (x, y) point on the 2D plane. They divide the 2D plane into a reasonably sized grid (e.g. 50x50), and for each grid cell they compute the average of all the activation vectors in the cell, visualize that activation vector using feature visualization, and put the resulting image into the grid cell. This gives a 50x50 grid of the "concepts" that the particular neural network layer we are analyzing can reason about. They also use attribution to show, for each grid cell, which class that grid cell most supports.

The paper then goes into a lot of detail about what we can infer from the activation atlas. For example, we can see that paths in activation vector space can correspond to human-interpretable concepts like the number of objects in an image, or moving from water to beaches to rocky cliffs. If we look at activation atlases for different layers, we can see that the later layers seem to get much more specific and complex, and formed of combinations of previous features (e.g. combining sand and water features to get a single sandbar feature).

By looking at images for specific classes, we can use attribution to see which parts of an activation atlas are most relevant for the class. By comparing across classes, we can see how the network makes decisions. For example, for fireboats vs. streetcars, the network looks for windows for both, crane-like structures for both (though less than windows), and water for fireboats vs. buildings for streetcars. This sort of analysis can also help us find mistakes in reasoning -- e.g. looking at the difference between grey whales and great white sharks, we can see that the network looks for the teeth and mouth of a great white shark, including an activation that looks suspiciously like a baseball. In fact, if you take a grey whale and put a patch of a baseball in the top left corner, this becomes an adversarial example that fools the network into thinking the grey whale is a great white shark. They run a bunch of experiments with these human-found adversarial examples and find they are quite effective.

Rohin's opinion: While the authors present this as a method for understanding how neurons interact, it seems to me that the key insight is about looking at and explaining the behavior of the neural network *on data points in-distribution*. Most possible inputs are off-distribution, and there is not much to be gained by understanding what the network does on these points. Techniques that aim to gain a global understanding of the network are going to be "explaining" the behavior of the network on such points as well, and so will be presenting data that we won't be able to interpret. By looking specifically at activations corresponding to in-distribution images, we can ensure that the data we're visualizing is in-distribution and is expected to make sense to us.

I'm pretty excited that interpretability techniques have gotten good enough that they allow us to construct adversarial examples "by hand" -- that seems like a clear demonstration that we are learning something real about the network. It feels like the next step would be to use interpretability techniques to enable us to actually fix the network -- though admittedly this would require us to also develop methods that allow humans to "tweak" networks, which doesn't really fit within interpretability research as normally defined.

Read more: OpenAI blog post and Google AI blog post

Feature Denoising for Improving Adversarial Robustness *(Cihang Xie et al)* (summarized by Dan H): This paper claims to obtain nontrivial adversarial robustness on ImageNet. Assuming an adversary can add perturbations of size 16/255 (l_infinity), previous adversarially trained classifiers could not obtain above 1% adversarial accuracy. Some groups have tried to break the model proposed in this paper, but so far it appears its robustness is close to what it claims, around 40% adversarial accuracy. Vanilla adversarial training is how they obtain said adversarial robustness. There has only been one previous public attempt at applying (multistep) adversarial training to ImageNet, as those at universities simply do not have the GPUs necessary to perform adversarial training on 224x224 images. Unlike the previous attempt, this paper ostensibly uses better hyperparameters, possibly accounting for the discrepancy. If true, this result reminds us that hyperparameter tuning can be critical even in vision, and that improving adversarial robustness on large-scale images may not be possible outside industry for many years.

Technical AI alignment

Learning human intent

Using Causal Analysis to Learn Specifications from Task Demonstrations *(Daniel Angelov et al)*

Reward learning theory

A theory of human values *(Stuart Armstrong)*: This post presents an outline of how to construct a theory of human values. First, we need to infer preferences and meta-preferences from humans who are in "reasonable" situations. Then we need to synthesize these into a utility function, by resolving contradictions between preferences, applying meta-preferences to preferences, and having a way of changing the procedures used to do the previous two things. We then need to argue that this leads to adequate outcomes -- he gives some simple arguments for this, that rely on particular facts about humans (such as the fact that they are scope insensitive).

Preventing bad behavior

Designing agent incentives to avoid side effects *(Victoria Krakovna et al)*: This blog post provides details about the recent update to the relative reachability paper (AN #10), which is now more a paper about the design choices available with impact measures. There are three main axes that they identify:

First, what baseline is impact measurede relative to? A natural choice is to compare against the starting state, but this will penalize the agent for environment effects, such as apples growing on trees. We can instead compare against an inaction baseline, i.e. measuring impact relative to what would have happened if the agent did nothing. Unfortunately, this leads to offsetting behavior: the agent first makes a change to get reward, and then undoes the change in order to not be penalized for impact. This motivates the stepwise inaction baseline, which compares each action against what would have happened if the agent did nothing *from that step onwards*.

Second, we need a measure by which to compare states. The unreachability measure measures how hard it is to reach the baseline from the current state. However, this "maxes out" as soon as the baseline is unreachability, and so there is no incentive to avoid further irreversible actions. This motivates relative reachability, which computes the set of states reachable from the baseline, and measures what proportion of those states are reachable from the state created by the agent. Attainable utility (AN #25) generalizes this to talk about the *utility* that could be achieved from the baseline for a wide range of utility functions. (This is equivalent to relative reachability when the utility functions are of the form "1 if state s is ever encountered, else 0".)

Finally, we need to figure how to penalize changes in our chosen measure. Penalizing decreases in the measure allows us to penalize actions that make it harder to do things (what the AUP post calls "opportunity cost"), while penalizing increases in the measure allows us to penalize convergent instrumental subgoals (which almost by definition increase the ability to satisfy many different goals or reach many different states).

Rohin's opinion: Since the AUP post was published about half a year ago, I've been watching this unification of AUP and relative reachability slowly take form, since they were phrased very differently initially. I'm glad to see this finally explained clearly and concisely, with experiments showing the effect of each choice. I do want to put special emphasis on the insight of AUP that the pursuit of convergent instrumental subgoals leads to large *increases* in "ability to do things", and thus that penalizing increases can help avoid such subgoals. This point doesn't typically make it into the academic writings on the subject but seems quite important.

On the topic of impact measures, I'll repeat what I've said before: I think that it's hard to satisfy the conjunction of three desiderata -- objectivity (no dependence on human values), safety (preventing any catastrophic outcomes) and usefulness (the AI system is still able to do useful things). Impact measures are very clearly aiming for the first two criteria, but usually don't have much to say about the third one. My expectation is that there is a strong tradeoff between the first two criteria and the third one, and impact measures have not dealt with this fact yet, but will have to at some point.

Conservative Agency via Attainable Utility Preservation *(Alexander Matt Turner et al)*: This paper presents in a more academic format a lot of the content that Alex has published about attainable utility preservation, see Towards a New Impact Measure (AN #25) and Penalizing Impact via Attainable Utility Preservation(AN #39).

Interpretability

Exploring Neural Networks with Activation Atlases *(Shan Carter et al)*: Summarized in the highlights!

Adversarial examples

Feature Denoising for Improving Adversarial Robustness *(Cihang Xie et al)*: Summarized in the highlights!

Forecasting

Signup form for AI Metaculus *(Jacob Lagerros and Ben Goldhaber)*: Recently, forecasting platform Metaculus launched a new instance dedicated specifically to AI in order to get good answers for empirical questions (such as AGI timelines) that can help avoid situations like info-cascades. While most questions don't have that many predictions, the current set of beta-users were invited based on forecasting track-record and AI domain-expertise, so the signal of the average forecast should be high.

Some interesting predictions include:

- By end of 2019, will there be an agent at least as good as AlphaStar using non-controversial, human-like APM restrictions? *[mean: 58%, median: 66%, n = 26]*

- When will there be a superhuman Starcraft II agent with no domain-specific hardcoded knowledge, trained using <=$10,000 of publicly available compute? *[50%: 2021 to 2037, with median 2026, n = 35]*

This forecast is supported by a Guesstimate model, which estimates current and future sample efficiency of Starcraft II algorithms, based on current performance, algorithmic progress, and the generalization of Moore's law. For algorithmic progress, they look at the improvement in sample efficiency on Atari, and find a doubling time of roughly a year, via DQN --> DDQN --> Dueling DDQN --> Prioritized DDQN --> PPO --> Rainbow --> IMPALA.

Overall, there are 50+ questions, including on malicious use of AI, publishing norms, conference attendance, MIRI's research progress, the max compute doubling trend, OpenAI LP, nationalisation of AI labs, whether financial markets expect AGI, and more. You can sign-up to join here.

AI conference attendance *(Katja Grace)*: This post presents data on attendance numbers at AI conferences. The main result: "total large conference participation has grown by a factor 3.76 between 2011 and 2019, which is equivalent to a factor of 1.21 per year during that period". Looking at the graph, it seems to me that the exponential growth started in 2013, which would mean a slightly higher factor of around 1.3 per year. This would also make sense given that the current boom is often attributed to the publication of AlexNet in 2012.

Field building

Alignment Research Field Guide *(Abram Demski)*: This post gives advice on how to get started on technical research, in particular by starting a local MIRIx research group.

Rohin's opinion: I strongly recommend this post to anyone looking to get into research -- it's a great post; I'm not summarizing it because I want this newsletter to be primarily about technical research. Even if you are not planning to do the type of research that MIRI does, I think this post presents a very different perspective on how to do research compared to the mainstream view in academia. Note though that this is *not* the advice I'd give to someone trying to publish papers or break into academia. Also, while I'm talking about recommendations on how to do research, let me also recommend Research as a Stochastic Decision Process.

Miscellaneous (Alignment)

Partial preferences needed; partial preferences sufficient *(Stuart Armstrong)*: I'm not sure I fully understand this post, but my understanding is that it is saying that alignment proposals must rely on some information about human preferences. Proposals like impact measures and corrigibility try to formalize a property that will lead to good outcomes; but any such formalization will be denoting some policies as safe and some as dangerous, and there will always exist a utility function according to which the "safe" policies are catastrophic. Thus, you need to also define a utility function (or a class of them?) that safety is computed with respect to; and designing this is particularly difficult.

Rohin's opinion: This seems very similar to the problem I have with impact measures, but I wouldn't apply that argument to corrigibility. I think the difference might be that I'm thinking of "natural" things that agents might want, whereas Stuart is considering the entire space of possible utility functions. I'm not sure what drives this difference.

Understanding Agent Incentives with Causal Influence Diagrams *(Tom Everitt et al)*: This post and associated paper model an agent's decision process using a causal influence diagram -- think of a Bayes net, and then imagine that you add nodes corresponding to actions and utilities. A major benefit of Bayes nets is that the criterion of d-separation can be used to determine whether two nodes are conditionally independent. Once we add actions and utilities, we can also analyze whether observing or intervening on nodes would lead the agent to achieve higher expected utility. The authors derive criteria resembling d-separation for identifying each of these cases, which they call observation incentives (for nodes whose value the agent would like to know) and intervention incentives (for nodes whose value the agent would like to change). They use observation incentives to show how to analyze whether a particular decision is fair or not (that is, whether it depended on a sensitive feature that should not be used, like gender). Intervention incentives are used to establish the security of counterfactual oracles more simply and rigorously.

Rohin's opinion: These criteria are theoretically quite nice, but I'm not sure how they relate to the broader picture. Is the hope that we will be able to elicit the causal influence diagram an AI system is using, or something like it? Or perhaps that we will be able to create a causal influence diagram of the environment, and these criteria can tell us which nodes we should be particularly interested in? Maybe the goal was simply to understand agent incentives better, with the expectation that more knowledge would help in some as-yet-unknown way? None of these seem very compelling to me, but the authors might have something in mind I haven't thought of.

Other progress in AI

Exploration

World Discovery Models *(Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo Avila Pires et al)*

Reinforcement learning

Learning Dynamics Model in Reinforcement Learning by Incorporating the Long Term Future *(Nan Rosemary Ke et al)*

Deep learning

Self-Tuning Networks: Bilevel Optimization of Hyperparameters using Structured Best-Response Functions *(Matthew MacKay, Paul Vicol et al)*

Hierarchical RL

Model Primitive Hierarchical Lifelong Reinforcement Learning *(Bohan Wu et al)*

Miscellaneous (AI)

The Bitter Lesson *(Rich Sutton)*: This blog post is controversial. This is a combination summary and opinion, and so is more biased than my summaries usually are.

Much research in AI has been about embedding human knowledge in AI systems, in order to use the current limited amount of compute to achieve some outcomes. That is, we try to get our AI systems to think the way we think we think. However, this usually results in systems that work currently, but then cannot leverage the increasing computation that will be available. The bitter lesson is that methods like search and learning that can scale to more computation eventually win out, as more computation becomes available. There are many examples that will likely be familiar to readers of this newsletter, such as chess (large scale tree search), Go (large scale self play), image classification (CNNs), and speech recognition (Hidden Markov Models in the 70s, and now deep learning).

Shimon Whiteson's take is that in reality lots of human knowledge has been important in getting AI to do things; such as the invariances built into convolutional nets, or the MCTS and self-play algorithm underlying AlphaZero. I don't see this as opposed to Rich Sutton's point -- it seems to me that the takeaway is that we should aim to build algorithms that will be able to leverage large amounts of compute, but we can be clever and embed important knowledge in such algorithms. I think this criterion would have predicted ex-ante (i.e. before seeing the results) that much past and current research in AI was misguided, without also predicting that any of the major advances (like CNNs) were misguided.

It's worth noting that this is coming from a perspective of aiming for the most general possible capabilities for AI systems. If your goal is to instead build something that works reliably now, then it really is a good idea to embed human domain knowledge, as it does lead to a performance improvement -- you should just expect that in time the system will be replaced with a better performing system with less embedded human knowledge.

One disagreement I have is that this post doesn't acknowledge the importance of data. The AI advances we see now are ones where the data has been around for a long time (or you use simulation to get the data), and someone finally put in enough engineering effort + compute to get the data out and put it in a big enough model. That is, currently compute is increasing much faster (AN #7) than data, so the breakthroughs you see are in domains where the bottleneck was compute and not data; that doesn't mean data bottlenecks don't exist.

News

AI Safety workshop at IJCAI 2019 *(HuÃ¡scar Espinoza et al)*: There will be a workshop on AI safety at IJCAI 2019 in Macao, China; the paper submission deadline is April 12. In addition to the standard submissions (technical papers, proposals for technical talks, and position papers), they are seeking papers for their "AI safety landscape" initiative, which aims to build a single document identifying the core knowledge and needs of the AI safety community.

Verifying vNM-rationality requires an ontology

Result
It is impossible to verify that an agent is vNM-rational by observing its actions without access to the domain of its utility function.

Motivation
Alphonso and Beatriz both go the market to buy fruit.
Alphonso prefers grapes to oranges.
He fills his basket with grapes and pays for them.

Beatriz carefully picks through the fruit and purchases some oranges and some grapes.
Callisto arrives with a package of grapes.

"Say, Beatriz, would you like to trade some of your oranges for this package of grapes?" Callisto offers.

"Gladly." Beatriz replies, exchanging some of her oranges for the grapes.

A few moments later, Alphonso notices Beatriz giving Deion some grapes in exchange for some oranges.

"You are acting irrationally, Beatriz!" Alphonso exclaims. "Your unstable preference between oranges and grapes makes it possible for a malicious agent to exploit you and exhaust your entire grocery budget!

"Ah, but I am acting rationally." Beatriz replied with a smile. "I prefer fruit that is fresh enough to last more than seven days. Thus, I trade away fruit that will spoil before that time."

Explanation
Consider an agent A.
We are interested in verifying whether or not A is vNM-rational.
However, we are only able to observe A's decisions without any access to the domain of A's utility function.

Without this access, it is impossible to distinguish between vNM-irrational choices (i.e. choices that violate one of the axioms of vNM-rationality) and choices that are vNM-rational but made under an unexpected ontology.

In other words, we need to know how A perceives outcomes of the world before we can verify that A's preferences over those outcomes are vNM-rational.

Book Review: The Eureka Factor

Cross posted from my personal blog.

Last month I finally got round to reading *The Eureka Factor* by John Kounios and Mark Beeman, a popular book summarising research on 'insightful' thinking. I first mentioned it a couple of years ago after I'd read a short summary article, when I realised it was directly relevant to my recurring 'two types of mathematician' obsession:

> The book is not focussed on maths – it’s a general interest book about problem solving and creativity in any domain. But it looks like it has a very similar way of splitting problem solvers into two groups, ‘insightfuls’ and ‘analysts’. ‘Analysts’ follow a linear, methodical approach to work through a problem step by step. Importantly, they also have cognitive access to those steps – if they’re asked what they did to solve the problem, they can reconstruct the argument.
>
> ‘Insightfuls’ have no such access to the way they solved the problem. Instead, a solution just ‘pops into their heads’.
>
> Of course, nobody is really a pure ‘insightful’ or ‘analyst’. And most significant problems demand a mixed strategy. But it does seem like many people have a tendency towards one or the other.

I wasn't too sure what I was getting into. The replication crisis has made me hyperaware of the dangers of uncritically accepting any results in psychology, and I'm way too ignorant of the field to have a good sense for which results still look plausible. However, the book turned out to be so extraordinarily Relevant To My Interests that I couldn't resist writing up a review anyway.

The final chapters had a few examples along the lines of '[weak environmental effect] primes people to be more/less insightful', and I know enough to stay away from those, but the earlier parts look somewhat more solid to me. I haven't made much effort to trace back references, though, and I could easily still be being too credulous.

(I didn't worry so much about replication with my previous post on the Cognitive Reflection Test. Getting the bat and ball question wrong is hardly the kind of weak effect that you need a sensitive statistical instrument to detect. It's almost impossible to *stop* people getting it wrong! I did steer clear of any more dubious priming-style results, though, like the claim that people do better on the CRT when reading it 'in a disfluent font'.)

Insight and intuition

First, it's worth getting clear on exactly what Kounious and Beeman mean by 'insight'. As they use it, insight is a specific type of creative thinking, which they define more generally as 'the ability to reinterpret something by breaking it down into its elements and recombining these elements in a surprising way to achieve some goal.' Insight is distinguished by its suddenness and lack of conscious control:

> When this kind of creative recombination takes place in an instant, it’s an insight. But recombination can also result from the more gradual, conscious process that cognitive psychologists call “analytic” thought. This involves methodically and

> deliberately considering many possibilities until you find the solution. For example, when you’re playing a game of Scrabble, you must construct words from sets of letters. When you look at the set of letters “A-E-H-I-P-N-Y-P” and suddenly realize that they can form the word “EPIPHANY,” then that would be an insight. When you systematically try out different combinations of the letters until you find the word, that’s analysis.

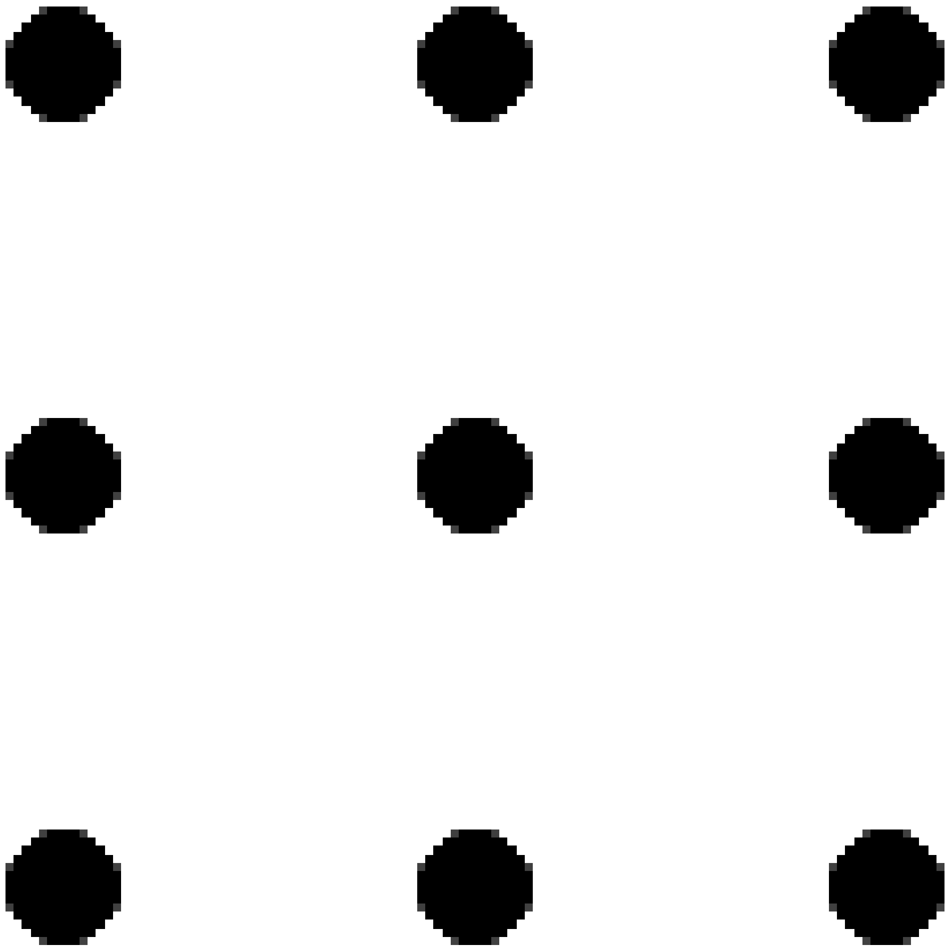
Insights tend to have a few other features in common. Solving a problem by insight is normally very satisfying: the insight comes into consciousness along with a small jolt of positive affect. The insight itself is usually preceded by a longer period of more effortful thought about the problem. Sometimes this takes place just before the moment of insight, while at other times there is an 'incubation' phase, where the solution pops into your head while you've taken a break from deliberately thinking about it.

I'm not really going to get into this part in my review, but the related word 'intuition' is also used in an interestingly specific sense in the book, to describe the sense that a new idea is lurking beneath the surface, but is not consciously accessible yet. Intuitions often precede an insight, but have a different feel to the insight itself:

> This puzzling phenomenon has a strange subjective quality. It feels like an idea is about to burst into your consciousness, almost as though you’re about to sneeze. Cognitive psychologists call this experience “intuition,” meaning an awareness of the presence of information in the unconscious mind — a new idea, solution, or perspective — without awareness of the information itself, at least until it pops into consciousness.

Insight problems

To study insight, psychologists need to come up with problems that reliably trigger an insight solution. One classic example discussed in *The Eureka Factor* is the Nine Dot Problem, where you are asked to connect the following 3 by 3 grid of black dots using only four lines, without retracing or taking your pen off the page:



If you've somehow avoided seeing this puzzle before, think about it for a while first. I've put the solution and my discussion of it in a spoiler block below:

A solution can be found in the Wikipedia article on insight problems here. It'll probably look irritatingly obvious once you see it. The key feature of the solution is that the lines you draw have to extend outside the confines of the square of dots you start with (thus spawning a whole subgenre of annoying business literature on 'thinking outside the box'). Nothing in the rules forbids this, but the setup focusses most people's attention on the grid itself, and breaking out of this mindset requires a kind of reframing, a throwing away of artificially imposed constraints. This is a common characteristic of insight problems.

This characteristic also makes insight hard to test. For testing purposes, it's useful to have a large stock of similar puzzles in hand. But a good reframing like the one in the

Nine Dot Problem tends to be a bit of a one-off: once you've had the idea of extending the lines outside the box, it applies trivially to all similar puzzles, and not at all to other types of puzzle.

(I talked about something similar in my last post, on the Cognitive Reflection Test. The test was inspired by one good puzzle, the 'bat and ball problem', and adds two other questions that were apparently picked to be similar. Five thousand words and many comments later, it's not obvious to me or most of the other commenters that these three problems form any kind of natural set at all.)

Kounios and Beeman discuss several of these eyecatching 'one-off' problems in the book, but their own research that they discuss is focussed on a more standardisable kind of puzzle, the Remote Associates Test. This test gives you three words, such as

PINE CRAB SAUCE

and asks you to find the common word that links them. The authors claim that these can be solved either with or without insight, and asked participants to self-categorise their responses as either fitting in the 'insightful' or 'analytic' categories:

> The analytic approach is to consciously search through the possibilities and try out potential answers. For example, start with “pine.” Imagine yourself thinking: What goes with “pine”? Perhaps “tree”? “Pine tree” works. “Crab tree”? Hmmm … maybe. “Tree sauce”? No. Have to try something else. How about “cake”? “Crab cake” works. “Cake sauce” is a bit of a reach but might be acceptable. However, “pine cake” and “cake pine” definitely don’t work. What else? How about “crabgrass”? That works. But “pine grass”? Not sure. Perhaps there is such a thing. But “sauce grass” and “grass sauce” are definitely out. What else goes with “sauce”? How about “applesauce”? That’s good. “Pineapple” and “crab apple” also work. The answer is “apple”!
>
> This is analytical thought: a deliberate, methodical, conscious search through the possible combinations. But this isn’t the only way to come up with the solution. Perhaps you’re trying out possibilities and get stuck or even draw a blank. And then, “Aha! Apple” suddenly pops into your awareness. That’s what would happen if you solved the problem by insight. The solution just occurs to you and doesn’t seem to be a direct product of your ongoing stream of thought.

This categorisation seems suspiciously neat, and if I rely on my own introspection for solving one of these (which is obviously dubious itself) it feels like more of a mix. I'll often generate some verbal noise about cakes and trees that sounds vaguely like I'm doing something systematic, but the main business of solving the thing seems to be going on nonverbally elsewhere. But I do think there's *something* there – the answer can be very immediate and 'poppy', or it can surface after a longer and more accessible process of trying plausible words. This was tested in a more objective way by seeing what people do when they *don't* come up with the answer:

> Insightfuls made more “errors of omission.” When waiting for an insight that hadn’t yet arrived, they had nothing to offer in its place. So when the insight didn’t arrive in time, they let the clock run out without having made a guess. In contrast, Analysts made more “errors of commission.” They rarely timed out, but instead guessed – sometimes correctly – by offering the potential solution they had been consciously thinking about when their time was almost up.

Kounios and Beeman's research focussed on finding neural correlates of the 'aha' moment of insight, using a combination of an EEG test to pinpoint the time of the insight, and fMRI scanning to locate the brain region:

> We found that at the moment a solution pops into someone’s awareness as an insight, a sudden burst of high-frequency EEG activity known as “gamma waves” can be picked up by electrodes just above the right ear. (Gamma waves represent cognitive processing in the brain, such as paying attention to something or linking together different pieces of information.) We were amazed at the abruptness of this burst of activity—just what one would expect from a sudden insight. Functional magnetic resonance imaging showed a corresponding increase in blood flow under these electrodes in a part of the brain’s right temporal lobe called the “anterior superior temporal gyrus” (see figure 5.2), an area that is involved in making connections between distantly related ideas, as in jokes and metaphors. This activity was absent for analytic solutions.
>
> So we had found a neural signature of the aha moment: a burst of activity in the brain’s right hemisphere.

I'm not sure how settled this is, though. I haven't tried to do a proper search of the literature, but certainly a review from 2010 describes the situation as very much in flux:

> A recent surge of interest into the neural underpinnings of creative behavior has produced a banquet of data that is tantalizing but, considered as a whole, deeply self-contradictory.

(The book was published somewhat later, in 2015, but mostly cites research from prior to this review, such as this paper.)

As an outsider it's going to be pretty hard for me to judge this without spending a *lot* more time than I really want to right now. However, regardless of how this holds up, I was really interested in the authors' discussion of *why* a right-hemisphere neural correlate of insight would make sense.

Insight and context

One of the authors, Mark Beeman, had previously studied language deficits in people who had suffered brain damage to the right hemisphere. One such patient was the trial attorney D.B.:

> What made D.B. “lucky” was that the stroke had damaged his right hemisphere rather than his left. Had the stroke occurred in the mirror-image left-hemisphere region, he would have experienced Wernicke’s aphasia, a profound deficit of language comprehension. In the worst cases, people with Wernicke’s aphasia may be completely unable to understand written or spoken language.
>
> Nevertheless, D.B. didn’t feel lucky. He may have been better off than if he’d had a left-hemisphere stroke, but he felt that his language ability was far from normal. He said that he “couldn’t keep up” with conversations or stories the way he used to. He felt impaired enough that he had stopped litigating trials—he thought that it would have been a disservice to his clients to continue to represent them in court.

> D.B. and the other patients were able to understand the straightforward meanings of words and the literal meanings of sentences. Even so, they complained about vague difficulties with language. They failed to grasp the gist of stories or were unable to follow multiple-character or multiple-plot stories, movies, or television shows. Many didn’t get jokes. Sarcasm and irony left them blank with incomprehension. They could sometimes muddle along without these abilities, but whenever things became subtle or implicit, they were lost.

An example of the kind of problem D.B. struggled with is the following:

> Saturday, Joan went to the park by the lake. She was walking barefoot in the shallow water, not knowing that there was glass nearby. Suddenly, she grabbed her foot in pain and called for help, and the lifeguard came running.

If D.B. was given a statement about something that occurred explicitly in the text, such as 'Joan went to the park on Saturday', he could say whether it was true or false with no problems at all. In fact, he did better than all of the control subjects on these sorts of explicit questions. But if he was instead presented with a statement like 'Joan cut her foot', where some of the facts are left implicit, he was unable to answer.

This was interesting to me, because it seems so directly relevant to the discussion last year on 'cognitive decoupling'. This is a term I'd picked up from Sarah Constantin, who herself got it from Keith Stanovich:

> Stanovich talks about “cognitive decoupling”, the ability to block out context and experiential knowledge and just follow formal rules, as a main component of both performance on intelligence tests and performance on the cognitive bias tests that correlate with intelligence. Cognitive decoupling is the opposite of holistic thinking. It’s the ability to separate, to view things in the abstract, to play devil’s advocate.

The patients in Beeman's study have so much difficulty with contextualisation that they struggle with anything at all that is left implicit, even straightforward inferences like 'Joan cut her foot'. This appears to match with other evidence from visual half-field studies, where subjects are presented with words on either the right or left half of the visual field. Those on the left half will go first to the right hemisphere, so that the right hemisphere gets a head start on interpreting the stimulus. This shows a similar difference between hemispheres:

> The left hemisphere is sharp, focused, and discriminating. When a word is presented to the left hemisphere, the meaning of that word is activated along with the meanings of a few closely related words. For example, when the word “table” is presented to the left hemisphere, this might strongly energize the concepts “chair” and “kitchen,” the usual suspects, so to speak. In contrast, the right hemisphere is broad, fuzzy, and promiscuously inclusive. When “table” is presented to the right hemisphere, a larger number of remotely related words are weakly invoked. For example, “table” might activate distant associations such as “water” (for underground water table), “payment” (for paying under the table), “number” (for a table of numbers), and so forth.

Why would picking up on these weak associations be relevant to insight? The story seems to be that this tangle of secondary meanings - the 'Lovecraftian penumbra of monstrous shadow phalanges' - works to pull your attention away from the obvious interpretation you're stuck with, helping you to find a clever new reframing of the problem.

This makes a lot of sense to me as a rough outline. In my own experience at least, the kind of thinking that is likely to lead to an insight experience feels softer and more diffuse than the more 'analytic' kind, more a process of sort of rolling the ideas around gently in your head and seeing if something clicks than a really focussed investigation of the problem. 'Thinking too hard' tends to break the spell. This fits well with the idea that insights are triggered by activation of weak associations.

Final thoughts

There's a lot of other interesting material in the book about the rest of the insight process, including the incubation period leading up to an insight flash, and the phenomenon of 'intuitions', where you feel that an insight is on its way but you don't know what it is yet. I'll never get through this review if I try to cover all of that, so instead I'm going to finish up with a couple of weak associations of my own that got activated while reading the book.

I've been getting increasingly dissatisfied with the way dual process theories split cognition into a fast/automatic/intuitive 'System 1' and a slow/effortful/systematic 'System 2'. System 1 in particular has started to look to me like an amorphous grab bag of all kinds of things that would be better separated out.

The Eureka Factor has pushed this a little further, by bringing out a distinction between two things that normally get lumped under System 1 but are actually very different. One obvious type of System 1-ish behaviour is routine action, the way you go about tasks you have done many times before, like making a sandwich or walking to work. These kinds of activities require very little explicit thought and generally 'just happen' in response to cues in the environment.

The kind of 'insightful' thinking discussed in *The Eureka Factor* would also normally get classed under System 1: it's not very systematic and involves a fast, opaque process where the answer just pops into your head without much explanation. But it's also very different to routine action. It involves deliberately choosing to think about a *new* situation, rather than one you have seen many times before, and a successful insight gives you a qualitatively new kind of understanding. The insight flash itself is a very noticeable, enjoyable feature of your conscious attention, rather than the effortless, unexamined state of absorbed action.

This was pointed out to me once before by Sarah Constantin, in the comments section of her Distinctions in Types of Thought:

> You seem to be lumping “flashes of insight” in with “effortless flow-state”. I don’t think they’re the same. For one thing, inspiration generally comes in bursts, while flow-states can persist for a while (driving on a highway, playing the piano, etc.) Definitely, “flashes of insight” aren’t the same type of thought as “effortful attention” — insight feels easy, instant, and unforced. But they might be their own, unique category of thought. Still working out my ontology here.

I'd sort of had this at the back of my head since then, but the book has really brought out the distinction clearly. I'm sure these aren't the only types of thinking getting shoved into the System 1 category, and I get the sense that there's a lot more splitting out that I need to do.

I also thought about how the results in the book fit in with my perennial 'two types of mathematician' question. (This is a weird phenomenon I've noticed where a lot of mathematicians have written essays about how mathematicians can be divided into two groups; I've assembled a list of examples here.) 'Analytic' versus 'insightful' seems to be *one* of the distinctions between groups, at least. It seems relevant to Poincaré’s version, for instance:

> The one sort are above all preoccupied with logic; to read their works, one is tempted to believe they have advanced only step by step, after the manner of a Vauban who pushes on his trenches against the place besieged, leaving nothing to chance.
>
> The other sort are guided by intuition and at the first stroke make quick but sometimes precarious conquests, like bold cavalrymen of the advance guard.

In fact, Poincaré once also gave a striking description of an insight flash himself:

> Just at this time, I left Caen, where I was living, to go on a geologic excursion under the auspices of the School of Mines. The incidents of the travel made me forget my mathematical work. Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step, the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformations I had used to define the Fuchsian functions were identical with those of non-Euclidean geometry. I did not verify the idea; I should not have had time, as, upon taking my seat in the omnibus, I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience' sake, I verified the result at my leisure.

If the insight/analysis split is going to be relevant here, it would require that people favour either 'analytic' or 'insight' solutions as a general cognitive style, rather than switching between them freely depending on the problem. The authors do indeed claim that this is the case:

> Most people can, and to some extent do, use both of these approaches. A pure type probably doesn’t exist; each person falls somewhere on an analytic-insightful continuum. Yet many—perhaps most—people tend to gravitate toward one of these styles, finding their particular approach to be more comfortable or natural.

This is based on their own research where they recorded participant's self-report of whether they were using a 'insight' or 'analytic' approach to solve anagrams, and compared it with EEG recordings of their resting state. They found a number of differences, including more right-hemisphere activity in the 'insight' group, and lower levels of communication between the frontal lobe and other parts of the brain, indicating a more disorderly thinking style with less top-down control. This may suggest more freedom to allow weak associations between thoughts to have a crack at the problem, without being overruled by the dominant interpretation.

Again, and you're probably got very bored of this disclaimer, I have no idea how well the details of this will hold up. That's true for pretty much every specific detail in the book that I've discussed here. Still, the link between insight and weak associations makes a lot of sense to me, and the overall picture certainly triggered some useful reframings. That seems very appropriate for a book about insight.

'This Waifu Does Not Exist': 100,000 StyleGAN & GPT-2 samples

This is a linkpost for https://www.thiswaifudoesnotexist.net/index.html

Do you like bullet points?

I think more naturally in bullet points, and I (sometimes) like reading posts that are written in bullet style. (This website is one of my favorites, and is written entirely in bullets).

(Disclaimer, although I wrote this post in bullet points because it was cute, I don't think it's the best exemplar of them. Or rather, it's an example of using bullet points to do rough thinking, rather than an example of using them to illustrate a complex argument)

I like bullet points because:

- *It's easier to skim, and build up a high level understanding of a post's structure.* If you understand a concept you can skip it and move on, if you want to drill down and understand it better you can do so.
    - Relatedly, it *exposes your cruxes* more readily. You can pick out and refute points, in a way that can be harder with meandering prose.
- *It's easier to hash out early stage ideas.* When I'm first thinking about something, my brain is jumping around and forming connections, developing a model at multiple levels of resolution. Bullet lists make this easier to keep track of.
    - I like this for *other people's* posts as well, since it feels more playful, like I can be part of their early generation process. I think LessWrong would be better if more people wrote more unpolished things to get early feedback on them, and bullet lists are a nice way to signal that something is still in development.
- *Prose often adds unnecessary cruft.* In the transition from bullets-to-prose, posts can go 2x-3x as long (or, when I go to write a short bullet summary of something I wrote in prose, it turns out to be much shorter, and the prose mostly unnecessary)

I had assumed this was a common experience, and that it was in fact a weakness of humanity that we didn't have better, more comprehensive bullet-point tools.

But, alas, *Typical Mind Fallacy.* It turned out a couple people on the LessWrong team reacted very negatively to bullet points. Concerns include:

- It's easy to think you've communicated more clearly than you have, because you didn't bother writing the connecting words between paragraphs.
- They're *harder* to read straight through. If you **include bold words,** readers might not bother reading the non-bold words, and miss nuance.
- "I like numbered arguments, since that makes it easier to respond to individual points. But unnumbered bullet lists are just hard to parse."
    - *[Alas, the LessWrong website currently doesn't enable this very well because our Rich Editor's implementation of numbered lists was annoying]*
- "I dunno man it's just really hard to read. My brain keeps trying to collapse the bullets like they're code."

I asked a couple more people, and they said "I dunno, bullet points seem fine. Depends on the situation?"

So...

I am curious what the LessWrong userbase thinks about them overall. Raise your hand if you think bullet points are fine? Terrible? Great? Any particular types of posts you prefer reading bullet-style, and types of posts you think fare poorly if not written in prose?

Three ways that "Sufficiently optimized agents appear coherent" can be false

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

There has been a couple of recent posts suggesting that Eliezer Yudkowsky's Sufficiently optimized agents appear coherent thesis does not seem useful because it's vacuously true: one obvious way to formalize "coherent" implies that all agents can be considered coherent. In a previous comment, I suggested that we can formalize "coherent" in a different way to dodge this criticism. I believe there's reason to think that Eliezer never intended "Sufficiently optimized agents appear coherent" to have an airtight argument and be universally true. (The Arbital post contains a number of caveats, including "If there is a particular kind of optimization pressure that seems sufficient to produce a cognitively highly advanced agent, but which also seems sure to overlook some particular form of incoherence, then this would present a loophole in the overall argument and yield a route by which an advanced agent with that particular incoherence might be produced".) In this post, I suggest that considering the ways in which it could be false can be a useful way to frame some recent ideas in AI safety. (Note that this isn't intended to be an exhaustive list.)

Distributional shift

Even a very powerful optimization process cannot train or test an agent in every possible environment and for every possible scenario (by this I mean some sequence of inputs) that it might face, and some optimization processes may not care about many possible environments/scenarios. Given this, we can expect that if an agent faces a new environment/scenario that's very different from what is was optimized for, it may fail to behave coherently.

(Jessica Taylor made a related point in Modeling the capabilities of advanced AI systems as episodic reinforcement learning: "When the test episode is similar to training episodes (e.g. in an online learning context), we should expect trained policies to act like a rational agent maximizing its expected score in this test episode; otherwise, the policy that acts as a rational agent would get a higher expected test score than this one, and would therefore receive the highest training score.")

A caveat to this caveat is that if an agent is optimized for a broad enough range of environments/scenarios, it could become an explicit EU maximizer, and keep doing EU maximization even after facing a distributional shift. (In this case it may be highly unpredictable what the agent's utility function looks like outside the range that it was optimized for. Humans can be considered a good example of this.)

Optimize for low compute

Eric Drexler suggested that one way to keep AIs safe is to optimize them to use few computing resources. If computing resources are expensive, it will often be less costly to accept incoherent behavior than to expend computing resources to reduce such incoherence. (Eliezer noted that such incoherence would only be removed "given the option of eliminating it at a reasonable computational cost".)

A caveat to this is that the true economic costs for compute will continue to fall, eventually to very low levels, so this depends on people assigning artificially high costs to computing resources (which Eric suggests that they do). However assigning an optimization cost for compute that is equal to its economic cost would often produce a more competitive AI, and safety concerns may not be sufficient incentive for an AI designer (if they are mostly selfish) to choose otherwise (because the benefits of producing a more competitive AI are more easily internalized than the costs/risks). One can imagine that in a world where computing costs are very low in an economic sense, but everyone is treating compute as having high cost for the sake of safety, the first person to *not* do this would gain a huge competitive advantage.

The optimizing process wants the agent to remain incoherent

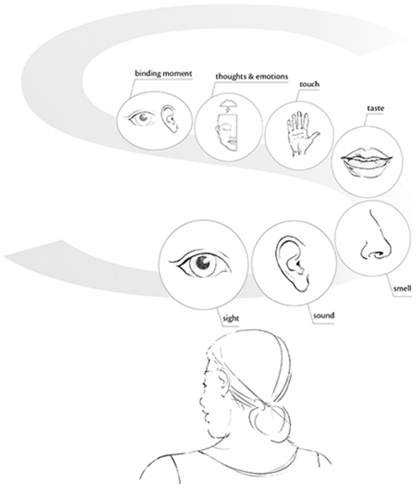
The optimizing process may itself be incoherent and not know how to become coherent or produce an agent that is coherent in an acceptable or safe way. A number of ideas fall into this category, including Peter Eckersley's recent Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function), which suggests that we should create AIs that handle moral uncertainty by randomly assigning a subagent (representing some moral theory) to each decision, with the argument that this is similar to how humans handle moral uncertainty. This can clearly be seen as an instance where the optimizing process (i.e., AI programmers) opts for the agent to remain incoherent because it does not know an acceptable/safe way to remove the incoherence.

A caveat here is that the agent may itself decide to become coherent anyway, and not necessarily in a way that the original optimizing process would endorse. For example, under Peter's proposal, one subagent may take an opportunity to modify the overall AI to become coherent in a way that it prefers, or multiple subagents may decide to cooperate and merge together into a more coherent agent. Another caveat is that incoherence is economically costly especially in a competitive multi-polar scenario, and if such costs are high enough the optimizing process may be forced to create a coherent agent even if it would prefer not to (in the absence of such costs).

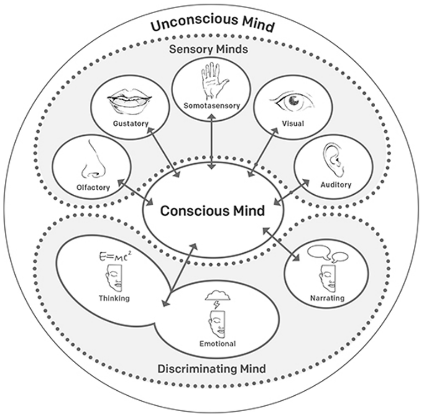
Subagents, introspective awareness, and blending

> 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. (If you didn't read my previous posts, this article has been written to also work as a stand-alone piece.)

The Mind Illuminated (Amazon, SSC review), or *TMI* for short, presents what it calls the *moments of consciousness model*. According to this model, our stream of consciousness consists of a series of discrete moments, each a mental object. Under this model, there are always different “subminds” which are projecting mental objects into consciousness. At different moments, different mental objects get selected as the content of consciousness.



If you’ve read some of the previous posts in this sequence, you may recognize this as sounding familiar. We started by discussing some of the neuroscience research on consciousness. There we covered the GWT/GNW theory of consciousness being a “workspace” in the brain that different brain systems project information into, and which allows them to synchronize their processing around a single piece of information. In the next post, we discussed the psychotherapy model of Internal Family Systems, which also conceives the mind of being composed of different parts, many of which are trying to accomplish various aims by competing to project various mental objects into consciousness. (TMI talks about subminds, IFS talks about parts, GWT/GNW just talks about different parts of the brain; for consistency’s sake, I will just use “subagent” in the rest of this post.)



At this point, we might want to look at some criticisms of this kind of a framework. Susan Blackmore has written an interesting paper called “There is no stream of consciousness”. She has several examples for why we should reject the notion of any stream of consciousness. For instance, this one:

> For many years now I have been getting my students to ask themselves, as many times as possible every day “Am I conscious now?”. Typically they find the task unexpectedly hard to do; and hard to remember to do. But when they do it, it has some very odd effects. First they often report that they always seem to be conscious when they ask the question but become less and less sure about whether they were conscious a moment before. With more practice they say that asking the question itself makes them more conscious, and that they can extend this consciousness from a few seconds to perhaps a minute or two. What does this say about consciousness the rest of the time?
>
> Just this starting exercise (we go on to various elaborations of it as the course progresses) begins to change many students’ assumptions about their own experience. In particular they become less sure that there are always contents in their stream of consciousness. How does it seem to you? It is worth deciding at the outset because this is what I am going to deny. I suggest that there is no stream of consciousness. […]
>
> I want to replace our familiar idea of a stream of consciousness with that of illusory backwards streams. At any time in the brain a whole lot of different things are going on. None of these is either ‘in’ or ‘out’ of consciousness, so we don’t need to explain the ‘difference’ between conscious and unconscious processing. Every so often something happens to create what seems to have been a stream. For example, we ask “Am I conscious now?”. At this point a retrospective story is concocted about what was in the stream of consciousness a moment before, together with a self who was apparently experiencing it. Of course there was neither a conscious self nor a stream, but it now seems as though there was. This process goes on all the time with new stories being concocted whenever required. At any time that we bother to look, or ask ourselves about it, it seems as though there is a stream of consciousness going on. When we don’t bother to ask, or to look, it doesn’t, but then we don’t notice so it doesn’t matter. This way the grand illusion is concocted.

This is an interesting argument. A similar example might be that when I first started doing something like track-back meditation when on walks, checking what was in my mind just a second ago. I was surprised at just how many thoughts I would have while on a walk, that I would usually just totally forget about afterwards, and come back home having no recollection of 95% of them. This seems similar to Blackmore's "was I conscious just now" question, in that when I started to check back the contents of my mind just a few seconds ago, I was frequently surprised by what I found out. (And yes, I've tried the "was I conscious just now" question as well, with similar results as Blackmore's students.)

Another example that Blackmore cites is change blindness. When people are shown an image, it's often possible to introduce unnoticed major changes into the image, as long as people are not looking at the very location of the change when it's made. Blackmore also interprets this as well to mean that there is no stream of consciousness - we aren't actually building up a detailed visual model of our environment, which we would then experience in our consciousness.

One might summarize this class of objections as something like, "stream-of-consciousness theories assume that there is a conscious stream of mental objects in our minds that we are aware of. However, upon investigation it often becomes apparent that we *haven't* been aware of something that was supposedly in our stream of consciousness. In change blindness experiments we weren't aware of what the changed detail actually was pre-change, and more generally we don't even have clear awareness of *whether we were conscious a moment ago*."

But on the other hand, as we reviewed earlier, there still seem to be objective experiments which establish the existence of something like a "consciousness", which holds approximately one mental object at a time.

So I would interpret Blackmore's findings differently. I agree that answers to questions like "am I conscious right now" are constructed somewhat on the spot, in response to the question. But I don't think that we need to reject having a stream of consciousness because of that. I think that *you can be aware of something, without being aware of the fact that you were aware of it*.

Robots again

For example, let's look at a robot that has something like a global consciousness workspace. Here are the contents of its consciousness on five successive timesteps:

1. It's raining outside
2. Battery low
3. Technological unemployment protestors are outside
4. Battery low
5. I'm now recharging my battery

Notice that at the first timestep, the robot was aware of the fact that it was raining outside; this was the fact being broadcast from consciousness to all subsystems. But at no later timestep was it conscious of the fact that at the first timestep, it had been aware of it raining outside. Assuming that no subagent happened to save this specific piece of information, then all knowledge of it was lost as soon as the content of the consciousness workspace changed.

But suppose that there is some subagent which happens to keep track of what has been happening in consciousness. In that case it may choose to make its memory of previous mind-states consciously available:

6. At time 1, there was the thought that [It’s raining outside]

Now there is a mental object in the robot’s consciousness, which encodes not only the observation of it raining outside before, *but also* the fact that the system was thinking of this before. That knowledge may then have further effects on the system - for example, when I became aware of how much time I spent on useless rumination while on walks, I got frustrated. And this seems to have contributed to making me ruminate less: as the system’s actions and their overall effect were metacognitively represented and made available for the system’s decision-making, this had the effect of the system adjusting its behavior to tune down activity that was deemed useless.

The Mind Illuminated calls this *introspective awareness.* Moments of introspective awareness are summaries of the system’s previous mental activity, with there being a dedicated subagent with the task of preparing and outputting such summaries. Usually it will only focus on tracking the specific kinds of mental states which seem important to track.

So if we ask ourselves “was I conscious just now” for the first time, that might cause the agent to output *some* representation of the previous state we were in. But it doesn’t have experience in answering this question, and if it’s anything like most human memory systems, it needs to have some kind of a concrete example of what exactly it is looking for. The first time we ask it, the subagent’s pattern-matcher knows that the system is presumably conscious at *this* instant, so it should be looking for some feature in our previous experiences which somehow resembles this moment, but it’s not quite sure of which one. And an introspective mind state, is likely to be different from the less introspective mind state that we were in a moment ago.

This has the result that on the first few times when it’s asked, the subagent may produce an uncertain answer: it’s basically asking its memory store “do my previous mind states resemble this one as judged by some unclear criteria”, which is obviously hard to answer.

With time, operating from the assumption that the system is currently conscious, the subagent may learn to find more connections between the current moment and past ones that it still happens to have in memory. Then it will report those as consciousness, and likely also focus more attention on aspects of the current experience which it has learned to consider “conscious”. This would match Blackmore’s report of “with more practice [the students] say that asking the question itself makes them more conscious, and that they can extend this consciousness from a few seconds to perhaps a minute or two”.

Similarly, this explains me not being aware of most of my thoughts, as well as change blindness. I had a stream of thoughts, but because I had not been practicing introspective awareness, there were no moments of introspective awareness making me aware of having had these thoughts. Though I was aware of the thoughts at the time, this was never re-presented in a way that would have left a memory trace.

In change blindness experiments, people might look at the same spot in a picture twice. Although they *did* see the contents of that spot at time 1 and were aware of them, that memory was never stored anywhere. When at time 2 they looked at the same spot and it was different, the lack of an awareness of what they saw previously means that they don’t notice the change.

Introspective awareness will be an important concept in my future posts. (Abram Demski also wrote a previous post on Track-Back Meditation, which is basically an exercise for introspective awareness.) Today, I’m going to talk about its relation to a concept I’ve talked about before: blending / cognitive fusion.

Blending

I’ve previously discussed “cognitive fusion”, as what happens when the content of a thought or emotion is experienced as an objective truth rather than a mental construct. For instance, you get angry at someone, and the emotion makes you experience them as a horrible person - and in the moment this seems just true to you, rather than being an interpretation created by your emotional reaction.

You can also fuse with more logical-type beliefs - or for that matter any beliefs - when you just treat them as unquestioned truths, without remembering the possibility that they might be wrong. In my previous post, I suggested that many forms of meditation were training the skill of intentional cognitive defusion, but I didn’t explain *how* exactly meditation lets you get better at defusion.

In my post about Internal Family Systems, I mentioned that IFS uses the term “blending” for when a subagent is sending emotional content to your consciousness, and suggested that IFS’s unblending techniques worked by associating extra content around those thoughts and emotions, allowing you to recognize them as mental objects. For instance, you might notice sensations in your body that were associated with the emotion, and let your mind generate a mental image of what the physical form of those sensations might look like. Then this set of emotions, thoughts, sensations, and visual images becomes “packaged together” in your mind, unambiguously designating it as a mental object.

My current model is that meditation works similarly, only using moments of introspective awareness as the “extra wrapper”. Suppose again that you are a robot, and the contents of your consciousness is:

1. It’s raining outside.

Note that this mental object is basically being taken as an axiomatic truth: what is in your consciousness, is that it is raining outside.

On the other hand, suppose that your consciousness contains this:

1. Sensor 62 is reporting that [it’s raining outside].

Now the mental object in your consciousness contains the origin of the belief that it’s raining. The information is made available to various subagents which have other beliefs. E.g. a subagent holding knowledge about sensors, might upon seeing this mental object, recognize the reference to “sensor 62” and output its estimate of that sensor’s reliability. The previous two mental objects could then be combined by a third subagent:

1. (Subagent A:) Sensor 62 is reporting that [it is raining outside]
2. (Subagent B:) Readings from sensor 62 are reliable 38% of the time.
3. (Subagent C:) It is raining outside with a 38% probability.

In my discussion of Consciousness and the Brain, I noted that one of the proposed functions of consciousness is to act as a production system, where many different subagents may identify particular mental objects and then apply various rules to transform the contents of consciousness as a result. What I've sketched above is exactly a sequence of production rules: at e.g. step 2, something like the rule "if sensor 62 is mentioned as an information source, output the current best estimate of sensor 62's reliability" is applied by subagent B. Then at the third timestep, another subagent combines the observations from the previous two timesteps, and sends *that* into consciousness.

What was important was that the system was not representing the outside weather just as an axiomatic statement, but rather it was explicitly representing it as a fallible piece of information with a particular source.

Here's something similar:

1. I am a bad person.
2. At t1, there was the thought that [I am a bad person].

Here, the moment of awareness is highlighting the nature of the previous thought as a thought, thus causing the system to treat it as such. If you used introspective awareness for unblending, it might go something like this:

1. *Blending:* you are experiencing everything that a subagent outputs as true. In this situation, there's no introspective awareness that would highlight those outputs as being just thoughts. "My friend is a horrible person", feels like a fact about the world.
2. *Partial blending:* you realize that the thoughts which you have might not be entirely true, but you still feel them emotionally and might end up acting accordingly. In this situation, there are moments of introspective awareness, but there are also enough of the original "unmarked" thoughts to also be affecting your other subagents. You feel hostile towards your friend, and realize that this may not be rationally warranted, but still end up talking in an angry tone and maybe saying things you shouldn't.
3. *Unblending:* all or nearly all of the thoughts coming from a subagent are being filtered through a mechanism that wraps them inside moments of introspective awareness, such as "this subagent is thinking that X". You know that you have a subagent which has this opinion, but none of the other subagents are treating it as a proven fact.

By training your mind to have more introspective moments of awareness, you will become capable of perceiving more and more mental objects as just that. A classic example would be all those mindfulness exercises where you stop identifying with the content of a thought, and see it as something separate from yourself. At more advanced levels, even mental objects which build up sensations such as those which make up the experience of a self may be seen as just constructed mental objects.

What I've Learned From My Parents' Arranged Marriage

When I tell people my parents had an arranged marriage, I get a number of different reactions. Most people have the wrong idea of exactly what that looks like, and those who do have the right idea often wonder if my parents can even understand what dating is like, given they've never experienced it. I've heard people assume that my parents' arranged marriage meant they were completely unable to help or give advice when it came to my dating life, and I've found the opposite to be the case; the advice my parents gave me about dating was as valuable as anything I found anywhere else, and allowed me to pass that advice on to my friends. Growing up hearing their story taught me a lot about what was important to know about myself before I started dating anyone, and how a good couple functions and grows together. I found that much of this is less commonly talked about when it comes to Western dating, and so I want to share their story and what I learned from it with you. For background, I'll start with telling you what arranged marriage is actually like.

Although some parts of India still do the traditional "bride and groom don't meet until the wedding", these tend to be remote and rural parts. Most arranged marriages today function a little more like a blind date, but with your parents and their network finding you a match rather than your friends. On the more traditional end, families may set up a "bride viewing", which today functions like a first meeting where the parents introduce each half of the couple, then leave them alone to get to know each other. They later tell their parents if they agree to the marriage or not. On the more liberal end, a couple may go on many dates before agreeing. In some cases, young people will date and fall in love, and the parents will meet after and decide to "arrange" the marriage if all parties agree to it. In the case of my parents, my dad's cousin (who he was very close with) met my mother and thought they would be a good match due to compatible philosophical interests and tastes in literature. My mother had, at that point, not dated at all, despite being in graduate school; it is normal for young people in India to feel marriage is not something they have to worry too much about because they trust their families will find someone good for them. The fact that my dad's cousin met my mother and immediately thought of my father points at another way arranged marriages affect the culture: people are always on the lookout for a good match.

When you ask someone who has had an arranged marriage about love, the first thing they say is that the love will come naturally once the couple is married. As a child, I always found this thought strange. As I grew older, though, I noticed the truth of this in the stories my mother told me about her relationship early on with my father. When they married, he was living in the US, and she was finishing her master's in India; for the year it took to finish her degree, they wrote letters. The way they did this nourished their love for each other, and fostered growth in their relationship. Western romance is described as something that happens on accident, but arranged romance happens on purpose. Even relationships that start with falling in love can benefit from growing and deepening that bond in the same way. This happens because you water love like a plant, and give it the right kinds of nutrients so it can grow.

One of the values that my mom spoke to me about more explicitly is that of cultural compatibility. In India, marriage is arranged through the social network of the parents. Traditionally, this focused a lot on social standing and religion, because of the idea that families of the same groups will raise their children similarly, and have similar values. My parents both grew up valuing learning and knowledge. They would have been far less compatible with people who were more focused on material wealth, or spiritual minimalism. Because their families had similar values, they were each instilled with similar values. This is reinforced by the fact that India is a more collectivist culture, and thus it is thought that your family knows you better than anyone else. Those who know you best are more likely to have a sense of who you would get along with, whether they're related to you or not. Further, getting along with the people your partner cares about most is important in any long term relationship. The fact that my mom got along well with my dad's cousin was a good sign; my mom connected more with the rest of my dad's family after the marriage, even though my dad had to go back to the US. Whether the relationship is arranged or not, fostering individual relationships with the people your partner cares about helps strengthen your relationship.

Compatibility includes not only what you value, but also what you want. Around the time my mother was getting married, many people her age were talking about wanting to move to the US. She was one of the few who wasn't fussed; she felt she'd be just as happy continuing to live in India. Of course, when she met my dad, that changed. For the right person, she was willing to move. There are people who wouldn't have been willing to make that move for anything, and there are those who wanted to move so badly that they didn't want to marry anyone willing to stay. This can be applied to anything one might want out of life, from living situation to religion to children, and more. In Western romantic media, this is often portrayed as being heartless. Ultimately, though, it's about trade-offs. Does your love for the person really overpower how much you want something? That answer differs for everyone. You can say that love conquers all, but a mismatch in this type of compatibility is one of the most common causes for divorce in the US. Knowing what you want your life to look like before you find the person to spend it with is going to be easier than trying to convince someone else to change what they want.

Of course, compatibility is nothing if you're not also complementary. This is where modern dating begins to look like marketing: know your target audience, and know what they want. If you know what kind of values you want your partner to have, you might already have a vague sense of what they would be like as a person. Knowing what you provide is crucial, especially when it comes to things like online dating. Traditional gender roles cover this well if you fit neatly in to one or the other, but things don't work that way for everyone. Give that my dad lived in the US, the fact that he could provide citizenship was huge. But he would not have been satisfied with a marriage with someone who saw this as his biggest asset. The fact that my mother was not obsessed with moving to the US meant that their complementary focus had to happen elsewhere. They shared the value of intellectual engagement, but my dad was always more focused on abstract ideas, while my mother tended to think more concretely. Here was where they were able to complement each other, which gave their life together more balance, and helped foster their growth individually as well. Finding someone whose traits and skills complement yours can help cover areas of life you struggle with, provide perspective when needed, and encourage you to grow and learn new things.

As a child, I didn't see the story of my parents as a love story. Love stories were about falling madly, hopelessly, and deeply, all at once, and my parents never really had that. But as I grew older, I noticed the details of their relationship. When my dad bought her a nice dress, it was as much because he wanted to see her in it as it was because he knew she hated shopping. When she challenged his ideas, it was out of love and respect, more than anything else. When we did things together as a family, they made sure to take time to connect with each other as a couple, even if it was only briefly. And as I became more independent, they were able to spend more and more time together. Love that lasts over a lifetime doesn't stay the same; it grows and changes with you as you grow and change. Falling in love doesn't happen once, but again and again.

Is there a difference between uncertainty over your utility function and uncertainty over outcomes?

I was discussing UDT yesterday and the question came up of how to treat uncertainty over your utility function. I suggested that this could be transformed into a question of uncertainty over outcomes. The intuition is that if you were to discover that apples were twice as valuable, you could simply pretend that you instead received twice as many apples. Is this approach correct? In particular, is it transformation compatible with UDT-style reasoning?

Alignment Newsletter #47

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

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

Highlights

AI Safety Needs Social Scientists *(Geoffrey Irving et al)* (summarized by Richard): One approach to AI safety is to "ask humans a large number of questions about what they want, train an ML model of their values, and optimize the AI system to do well according to the learned values". However, humans give answers that are limited, biased and often in disagreement with each other, and so AI safety needs social scientists to figure out how to improve this data - which eventually may be gathered from thousands or millions of people. Of particular importance is the ability to design rigorous experiments, drawing from an interdisciplinary understanding of human cognition and behaviour. The authors discuss Debate (AN #5) as a case study of a safety technique whose success depends on empirical questions such as: how skilled are humans as judges by default? Can we train people to be better judges? Are there ways to restrict debate to make it easier to judge?

There are a couple of key premises underlying this argument. The first is that, despite human biases, there are correct answers to questions about human values - perhaps defined as the answer we would endorse if given all relevant information and unlimited time to think. However, it’s not necessary for AIs to always find those answers, as long as they are able to recognise cases in which they’re uncertain and do nothing (while there are some cases in which inaction can cause harm, such as a self-driving car ceasing to steer mid-journey, it seems that the most worrying long-term catastrophes can be avoided by inaction). Another reason for optimism is that even incomplete or negative results from social science experiments may be useful in informing technical safety research going forward. However, in some cases the systems we're trying to reason about are very different from anything we can test now - for example, AI debaters that are much stronger than humans.

Richard's opinion: This post, and its accompanying paper, seems very sensible to me. While I have some doubts about how informative human debate data will be about superhuman debaters, it certainly seems worth trying to gain more empirical information. Note that while the paper primarily discusses Debate, I think that many of its arguments are applicable to any human-in-the-loop safety methods (and probably others too). Currently I think Ought is the safety group focusing most on collecting human data, but I look forward to seeing other researchers doing so.

Technical AI alignment

Technical agendas and prioritization

FLI Podcast: AI Breakthroughs and Challenges in 2018 with David Krueger and Roman Yampolskiy *(Ariel Conn, David Krueger and Roman Yampolskiy)*: David and Roman review AI progress in 2018 and speculate about its implications. Roman identified a pattern where we see breakthroughs like AlphaZero (AN #36), AlphaStar (AN #43) and AlphaFold (AN #36) so frequently now that it no longer seems as impressive when a new one comes out. David on the other hand sounded less impressed by progress on Dota and StarCraft, since both AI systems were capable of executing actions that humans could never do (fast reaction times for Dota and high actions-per-minute for StarCraft). He also thought that these projects didn't result in any clear general algorithmic insights the way AlphaZero did.

On the deep RL + robotics side, David identified major progress in Dactyl (AN #18) and QT-Opt (which I remember reading and liking but apparently I failed to put in the newsletter). He also cited GANs as having improved significantly, and talked about feature-wise transformations in particular. Roman noted the improving performance of evolutionary algorithms.

David also noted how a lot of results were obtained by creating algorithms that could scale, and then using a huge amount of compute for them, quoting AI and Compute (AN #7), Interpreting AI Compute Trends (AN #15) and Reinterpreting AI and Compute (AN #38).

On the policy side, they talked about deep fakes and the general trend that AI may be progressing to fast for us to keep up with its security implications. They do find it promising that researchers are beginning to accept that their research does have safety and security implications.

On the safety side, David noted that the main advance seemed to be with approaches using superhuman feedback, including debate (AN #5), iterated amplification (discussed frequently in this newsletter, but that paper was in AN #30) and recursive reward modeling (AN #34). He also identified unrestricted adversarial examples (AN #24) as an area to watch in the future.

Rohin's opinion: I broadly agree with the areas of AI progress identified here, though I would probably also throw in NLP, e.g. BERT. I disagree on the details -- for example, I think that OpenAI Five (AN #13) was much better than I would have expected at the time and the same would have been true of AlphaStar if I hadn't already seen OpenAI Five, and the fact that they did a few things that humans can't do barely diminishes the achievement at all. (My take is pretty similar to Alex Irpan's take in his post on AlphaStar.)

Treacherous Turn, Simulations and Brain-Computer Interfaces *(Michaël Trazzi)*

Learning human intent

AI Alignment Podcast: Human Cognition and the Nature of Intelligence *(Lucas Perry and Joshua Greene)* (summarized by Richard): Joshua Greene's lab has two research directions. The first is how we combine concepts to form thoughts: a process which allows us to understand arbitrary novel scenarios (even ones we don't think ever occurred). He discusses some of his recent reseach, which uses brain imaging to infer what's happening when humans think about compound concepts. While Joshua considers the combinatorial nature of thought to be important, he argues that to build AGI, it's necessary to start with "grounded cognition" in which representations are derived from perception and physical action, rather than just learning to manipulate symbols (like language).

Joshua also works on the psychology and neuroscience of morality. He discusses his recent work in which participants are prompted to consider Rawls' Veil of Ignorance argument (that when making decisions affecting many people, we should do so as if we don't know which one we are) and then asked to evaluate moral dilemmas such as trolley problems. Joshua argues that the concept of impartiality is at the core of morality, and that it pushes people towards more utilitarian ideas (although he wants to rebrand utilitarianism as "deep pragmatism" to address its PR problems).

Imitation Learning from Imperfect Demonstration *(Yueh-Hua Wu et al)*

Learning User Preferences via Reinforcement Learning with Spatial Interface Valuing *(Miguel Alonso Jr)*

Interpretability

Regularizing Black-box Models for Improved Interpretability *(Gregory Plumb et al)*

Robustness

Adversarial Examples Are a Natural Consequence of Test Error in Noise *(Nicolas Ford, Justin Gilmer et al)* (summarized by Dan H): While this was previously summarized in AN #32, this draft is much more readable.

Improving Robustness of Machine Translation with Synthetic Noise *(Vaibhav, Sumeet Singh, Craig Stewart et al)* (summarized by Dan H): By injecting noise (such as typos, word omission, slang) into the training set of a machine translation model, the authors are able to improve performance on naturally occurring data. While this trick usually does not work for computer vision models, it can work for NLP models.

Push the Student to Learn Right: Progressive Gradient Correcting by Meta-learner on Corrupted Labels *(Jun Shu et al)*

Miscellaneous (Alignment)

AI Safety Needs Social Scientists *(Geoffrey Irving et al)*: Summarized in the highlights!

AI strategy and policy

Humans Who Are Not Concentrating Are Not General Intelligences *(Sarah Constantin)*: This post argues that humans who skim the stories produced by GPT-2 (AN #46) would not be able to tell that they were generated by a machine, because while skimming we are not able to notice the obvious logical inconsistencies in its writing. Key quote: "OpenAI HAS achieved the ability to pass the Turing test against humans on autopilot". This suggests that fake news, social manipulation, etc. will become much easier. However, it might also force people to learn the skill of detecting the difference between humans and bots, which could let them learn to tell when they are actively focusing on something and are "actually learning" as opposed to skimming for "low order correlations".

Rohin's opinion: I noticed a variant of this effect myself while reading GPT-2 results -- my brain very quickly fell into the mode of skimming without absorbing anything, though it felt more like I had made the evaluation that there was nothing to gain from the content, which seems okay if the goal is to avoid fake news. I also find this to be particularly interesting evidence about the differences between our low-level, effortless pattern matching, as well as our more effortful and accurate "logical reasoning".

Other progress in AI

Exploration

InfoBot: Transfer and Exploration via the Information Bottleneck *(Anirudh Goyal et al)*

Reinforcement learning

An Overdue Post on AlphaStar *(Alex Irpan)*: The first post in this two-parter talks about the impact of AlphaStar (AN #43) on the StarCraft community and broader public. I'm focusing on the second one, which talks about AlphaStar's technical details and implications. Some of this post overlaps with my summary of AlphaStar, but those parts are better fleshed out and have more details.

First, imitation learning is a surprisingly good base policy, getting to the level of a Gold player. It's surprising because you might expect the DAgger problem to be extreme: since there are so many actions in a StarCraft game, your imitation learning policy will make some errors, and those errors will then compound over the very long remainder of the episode as they take the policy further away from normal human play into states that the policy wasn't trained on.

Second, population-based training is probably crucial and will be important in the future, because it allows for exploring the full strategy space.

Third, the major challenge is making RL achieve okay performance, and after that they very quickly become great. It took years of research to get Dota and StarCraft bots reach decent play, and then a few days of more training got them to be world class. Fun quote: "although OpenAI’s DotA 2 agent lost against a pro team, they were able to beat their old agent 80% of the time with 10 days of training".

Fourth, there were a lot of research results that went into AlphaStar. This suggests that there are large gains to be had by throwing a lot of techniques together and seeing how well they work, which doesn't happen very much currently. There are good reasons for this: it's much easier to evaluate a technique if its built upon a simple, standard algorithm rather than having to consider all of its interactions with other techniques which you may or may not be able to properly compare against. Still, there are going to be some cool results that we could do now if we just threw the right things together, and this sort of work also lets us test techniques in new settings to see which ones actually work in general, as opposed to only in the original evaluation.

Rohin's opinion: I really like this post, and agree with almost everything in it. On the imitation learning point, I also found it surprising how well imitation learning worked. Alex suggests that it could be that human data has enough variation that the agent can learn how to recover from incorrect decisions it could make. I think this is a partial explanation at best -- there is a huge combinatorial explosion, it's not clear why you don't need a much larger dataset to cover the entire space. Maybe there are "natural" representations in any realistic complex environment that you start to accurately learn at the level of compute that they're using, and once you have those then imitation learning with sufficient variation can work well.

On the last point about tossing techniques together, I think this might sometimes be worth doing but often may not be. It makes sense to do this with any real task, since that's a test of the technique against reality. (Here StarCraft counts as a "real" task while Atari does not; the criterion is something like "if the task is successfully automated we are impressed regardless of how it is solved".) I'm less keen on tossing techniques together for artificial benchmarks. I think typically these techniques improve the sample efficiency by a constant multiplicative factor by adding something akin to a good inductive bias; in that case throwing them together may let us solve the artificial benchmark sooner but it doesn't give us great evidence that the "inductive bias" will be good for realistic tasks. I think I don't actually disagree with Alex very much on the object-level recommendations, I would just frame them differently.

Learning to Generalize from Sparse and Underspecified Rewards *(Rishabh Agarwal et al)*

Reward Shaping via Meta-Learning *(Haosheng Zou, Tongzheng Ren et al)*

Investigating Generalisation in Continuous Deep Reinforcement Learning *(Chenyang Zhao et al)*

Deep learning

Random Search and Reproducibility for Neural Architecture Search *(Liam Li et al)*

News

MIRI Summer Fellows Program *(Colm Ó Riain)*: CFAR and MIRI are running the MIRI Summer Fellows Program from August 9-24. Applications are due March 31.

RAISE is launching their MVP *(Toon Alfrink)*: The Road to AI Safety Excellence will begin publishing lessons on inverse reinforcement learning and iterated amplification on Monday. They are looking for volunteers for their testing panel, who will study the material for about one full day per week, with guidance from RAISE, and provide feedback on the material and in particular on any sources of confusion.

How much funding and researchers were in AI, and AI Safety, in 2018?

I'm trying to build up a picture of how "much" research is going into general AI capabilities, and how much is going into AI safety.

The *ideal* question I'd be asking is "how much progress [measured in "important thoughts/ideas/tools" was being made that plausibly could lead to AGI in 2018, and how much progress was made that could plausibly lead to safe/aligned AI].

I assume that question is nigh impossible, so instead asking the approximation:

a) how much money went into AI capabilities research in 2018

b) how much money went into AI alignment research in 2018

c) how many researchers (ideally "research hours" but I'll take what I can get) were focused on capabilities research in 2018

d) how many researchers were focused on AI safety in 2018?

Blegg Mode

This is a linkpost for http://unremediatedgender.space/2018/Feb/blegg-mode/

Fanfiction for the blegg/rube parable in "A Human's Guide to Words", ~800 words. (*Content notice*: in addition to making a point about epistemology (which is why it may have been worth sharing here), this piece is also an obvious allegory about a potentially mindkilling topic; read with caution, as always.)

On the Regulation of Perception

Have you ever had one of those unexpected flashes of metacognition?

It was towards the end of long, successful day. Around 9:30PM.

I'd woken late so I'd be up a while longer, and was sort of lounging for a little while, sometimes reading a little, sometimes half-napping and dozing a bit on the couch.

Then, a simple enough thought: "I'm hungry. Perhaps, Arabic food."

Suddenly, a strange self-awareness and metacognition kicks in. Wholly unexpectedly, in fact. I wasn't in any sort of philosophical mood, wasn't analyzing, it just — there it was.

Wait. What is this "I - am - hungry"? What's going on here?

Followed by an unexpected and delightful rush of questions — and possibly some answers.

###

To just cut right to the chase — I'm afraid I'm failing to "show to my work," but we're all busy, eh? — it dawns on me that perhaps we're not regulating our bodies, but in fact regulating our subjective perception.

Hmm. How do I do this in a less wordy way?

I reckon the common perception is something like this —

"We eat because we're hungry."

But that's not precisely true. It's actually more like this —

"We eat because we ***perceive*** we're hungry."

"Hunger" isn't a single thing; it's a rollup of a lot of things.

You could be at low blood sugar or high ghrelin levels; your stomach could be physically empty; or perhaps blood sugar is high, ghrelin is low, your stomach is somewhat full, but you're mistaking boredom for hunger.

Regardless, the body isn't quite so precise as the gasoline tank on your car, which a well-calibrated gasoline gauge will tell you is approaching empty.

No, not so. Hunger is a rollup of a lot of factors, and it's a… perception/feeling/something. In any event, we can safely say it's subjective. Blood sugar levels, ghrelin levels, stomach contents — these are objective enough and could be measured. But you can't necessarily predict the subjective experience of hunger from them — two people could have identical levels of blood sugar, ghrelin, and stomach contents, and one could be hungry while the other is not.

###

Stepping back from hunger for a moment, let’s talk about ontology.

Ontology is a fancy word for “being.” Wikipedia kindly informs us that it involves questions like,

"*What can be said to exist? What is a thing? Into what categories, if any, can we sort existing things? What are the meanings of being? What are the various modes of being of entities?*"

So, there’s that.

I’m only interested in thing from ontology right now —

*“I **am** hungry.*"

I’m absolutely shocked that I didn’t notice this before, but I just realized that the default description of hunger is so *total and all-encompassing*.

###

Contrast —

“I perceive hunger.”

“I analyze and note reasons for hunger.”

“I decide to eat.”

###

We’ve traversed firmly into E-Prime territory at this point, one of my favorite tools for clarifying thought — albeit an expensive and awkward tool to deploy.

“E-Prime (short for English-Prime or English Prime) is a version of the English language that excludes all forms of the verb to be, including all conjugations, contractions and archaic forms."

So the words am, is, be, etc — gone, stricken from one’s vocabulary.

There’s offers some clarity in precise thinking, but writing in E-Prime often feels awkward; the word “is” makes it easier to string sentences together.

But the upside comes from making descriptions less total, absolutely, and all-encompassing —

"*Alfred Korzybski counseled his students to eliminate the infinitive and verb forms of* "*to be*" *from their vocabulary, whereas a second group continued to use* "*I am,*" "*You are,*" "*They are*" *statements as usual. For example, instead of saying,* "*I am depressed,*" *a student was asked to eliminate that emotionally primed verb and to say something else, such as,* "*I feel depressed when …*" *or* "*I tend to make myself depressed about …”*"

###

Back to hunger.

Re: hunger, there's no there there.

We perceive hunger rather than *are* hunger.

Generalizing a bit, among the daily chores for staying alive as a human include eating, drinking, sleeping, etc.

But we don't typically think or say, "I perceive hunger" and then to choose to eat, nor "I assess my adenosine levels are high" and then choose to sleep.

Nope.

Rather —

I am hungry.

I am tired.

###

The first sentence that occurred to me while lazing about with my book was, "We're not regulating our bodies — we're regulating our perception."

It's even more clear when looking at it from the opposite point of view — non-food substances that increase or decrease perceived hunger.

Ephedrine hydrochloride, especially when mixed with caffeine, is often used by bodybuilders to increase metabolic rate; i.e., it helps you burn fat by metabolizing it. Ephedrine gives you increased capacity for power generation and increased metabolization at the potential downside of increased thermogenesis (it makes you very warm - actually, this is useful in winter but a hassle otherwise) and it taxes one's heart when consumed, which is the largest downside.

Notably though, ephedrine is also an appetite suppressant. With the same blood sugar, ghrelin, stomach contents, etc etc — whilst taking ephedrine, you'll typically feel less hungry.

So you'll probably eat less.

The chain reaction from ephedrine does a lot of things biochemically, but it also *changes one's perception of hunger* — or the lack thereof*.*

###

I was feeling well-pleased with myself at this point, ready to plant a flag in the ground and and make triumphant declarations.

But then I think, "Well, hold on a minute, is there any time we consume food when we're not regulating perception?"

The first example of that might be some social setting — a case where you're not hungry, but it might be rude to decline food.

I thought about it more, and said, no, this is the same thing — it's consuming food in response to a perception that it'd be rude not to. Social cues and social settings are probably more complex modes of thought and perception, but it's still perceived.

“It would **be rude** of me not to eat here.”

I’d contend that, most of the time, that’s still regulating perception. Not the perception of hunger, but the perception of social decorum, or the perception of appropriate behavior, or something like that.

Still though, most of the time, regulating perception.

###

But then an actual counterexample came to mind — what about a person who puts together a detailed nutrition and meal plan, either to gain weight or lose weight, along with detailed recipes and meals to be consumed at certain times each day?

If someone pre-cooks a week’s worth of healthy food, and pre-commits to eating it at certain times, and to not skip those meals even if they’re not hungry, and to not have extra food that’s off the plan even if they’re hungry — this seems like something else entirely, does it not?

###

Someone more well-versed than me could weigh in on the relevant neuroscience — I’d imagine that detailed planning of nutritional consumption over time would include much more use of the neocortex than the limbic system.

But neuroscience and brain regions aside, it seems to me that moving off “I am hungry” offers some great possibility for better health, wellness, and control over one’s life.

The default inclination seems to be,

1. Some perception occurs to us. (Hunger, let’s say.)

2. Uncritically, we think, “I am hungry.”

3. We eat.

Whereas, with a bit of foresight, we might do this —

1. Do some research to figure out what nutrition and life patterns we’d like to live.

2. Get those plans to the right mix of optimal for health and sustainable.

3. Build those plans out so they’re easily runnable.

4. When later noting perceptions (ex, hunger), simply note them and then carry on with the plan as before.

Perhaps easier said than done!

At the very least, though, I think that replacing “I am hungry” with “I perceive myself to be hungry” has gotta offer some profit in terms of increased control and awareness.

Having a highly-imperative and all-encompassing nature of one's perceptions seems to be a major liability in the modern world.

So, perhaps a better title of this piece could have been, “On the Regulation of the Regulation of Perception”...

Nevertheless — interesting topic, eh?

Now, if you’ll excuse me, I have some shishtaowk, rice, hummous, and olive oil to attend to…

Asking for help teaching a critical thinking class.

I will soon be teaching a "critical thinking" class for undergraduates. Feel free to mentally replace "critical thinking" with "epistemic virtue". I would appreciate answers to any of these questions:

What would you do if you only had one three hour class period to teach a group of 15-30 undergrads "critical thinking"?

If you know me (Ronny Fernandez) what one educational objective would you give me, knowing it will not be my only one, to come up with a plan to achieve?

How would you given three hours or less make it so that 15-30 students are all absolutely sure of something, and then realize they were all wrong, without feeling like them feeling like you had cheated or done anything shady?

What are the most transferable rationality skills that are abundant in our community and rare elsewhere and how do you transfer them in a classroom setting?

How do you teach the things in the vicinity of scout mindset vs warrior mindset, arguments as soldiers, politics is the mind killer, without making your student temporarily dumber by giving them a fully general excuse to not engage with any disagreement they feel like ignoring?

How do you teach the difference between the kind of cognition you use to figure out how to get to your friends house and the kind of cognition Malfoy used in hpmor to think about the heritability of magic?

How do you teach the virtue of lightness, the virtue of curiosity?

How do you teach fallacies and cognitive biases without making your students temporarily dumber by giving them a fully general excuse to disregard any position or thinker they disagree with?

A Concrete Proposal for Adversarial IDA

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

Note: This post came out of a conversation with Geoffrey Irving and Buck Shlegeris.

Epistemic Status: I suspect Paul has already thought of most or all of the ideas presented here, though I nevertheless found the exercise of carefully specifying an IDA implementation helpful and suspect others may find reading it helpful as well.

This is a proposal for how to train a machine learning model to approximate HCH using Iterated Distillation and Amplification (IDA). This particular proposal came out of a desire to use a debate-like adversary to improve the amplification process, and the primary goal of this proposal is to show how one could do that. Though I have tried to retain a lot of the relevant detail, I have made two simplifications to make this proposal easier to specify: I am attempting to approximate something closer to weak HCH rather than strong HCH and I am only allowing the generation of two subquestions at a time. I am confident that those simplifications could easily be dropped, though I think doing so here would only make this presentation more complicated.

Before I proceed, I want to make one final note: this is not a proposal for how to build an aligned AGI. I think there are still a whole bunch of issues that would prevent this proposal from actually working.

Definitions

We will start with some initial definitions:

- Let Q be the set of all questions in natural language.
- Let A be the set of all answers in natural language.
- Let M be the sum type of either $Q \times Q$ or A representing either an answer to the given question or two subquestions to help answer it.
- Let $H : Q \to A$ be the answer that a human gives to the given question.
- Let $H_{\text{fan out}} : Q \to M$ be the answer or subquestion pair generated by a human when asked what to do with the given question.
- Let $H_{\text{fan in}} : Q \times (Q \times A) \times (Q \times A) \to M$ be the answer or two subquestions generated by a human to some question when given answers to two subquestions related to that question.
- Let $ML : Q \to \Delta(A)$ be a model (the training procedure for which we will describe below) from questions to a probability distribution over strings representing answers. Specifically, we will implement the probability distribution by having our model output an embedding vector which yields the probability distribution when fed into some trained language model (by repeatedly conditioning on previous characters and multiplying all the conditional probabilities).
- Let $ML_{\text{fan out}} : Q \to \Delta(M)$ be the function from questions to an embedding vector representing a distribution over answers or subquestions generated by asking ML what to do with the given question. For the two subquestion case we enforce that the distribution is symmetric wrt interchange of the subquestions.
- Let $ML_{\text{fan in}} : Q \times (Q \times A) \times (Q \times A) \to \Delta(M)$ be the function from two answers to an embedding vector representing a distribution over answers or subquestions generated by asking ML to integrate the given subquestion answers into an answer to the original question. We again enforce symmetry wrt interchange of the subquestions.
- Let $Adv : Q \times A \to \mathbb{R}$ be an adversary model which we will train (as described below) to predict how good the given answer is to the given question.
- Let $Adv_{\text{fan out}} : Q \times M \to \mathbb{R}$ be an adversary model for $ML_{\text{fan out}}$ generated by calling Adv using the same transformation as used to implement $ML_{\text{fan out}}$ from ML.

- Let $Adv_{\text{fan in}} : Q \times (Q \times A) \times (Q \times A) \times M \to R$ be an adversary model for $ML_{\text{fan in}}$ generated by calling Adv using the same transformation as used to implement $ML_{\text{fan in}}$ from ML.
- Let ϵ be some Gumbel random variable such that each use of ϵ below is its own independent sample from a Gumbel distribution. We use a Gumbel random variable so that we can use the Gumbel-max trick to effectively do Boltzmann exploration.

Amplification

The goal here is to use the adversary to determine which portions of the HCH tree we actually go down to maximize our sample efficiency wrt H. Specifically, the amplification procedure is as follows:

1. Generate some question $Q \in Q$ either from the training distribution or from the output of some adversary.

2. Call $H_{\text{fan out}}(Q)$. If we get an answer, let $A = H_{\text{fan out}}(Q)$ and skip to the end. Otherwise, let $SQ_1, SQ_2 = H_{\text{fan out}}(Q)$.

3. Call Adv to generate scores $S_1 = Adv(SQ_1, ML(SQ_1))$ and $S_2 = Adv(SQ_2, ML(SQ_2))$.

4. Without loss of generality, let $S_1 + \epsilon \leq S_2 + \epsilon$ and recurse on SQ_1 (by setting $Q = SQ_1$ in the subcall) to generate A_1 (from the A of the subcall) and let $A_2 = ML(SQ_2)$.

5. Call $H_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$. If we get an answer, let $A = H_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$. Otherwise, let $SQ_1, SQ_2 = H_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$ and return to step 3.

This procedure should eventually produce an answer, though we then need to compute the following losses for $ML, ML_{\text{fan out}}, ML_{\text{fan in}}, Adv, Adv_{\text{fan out}}, Adv_{\text{fan in}}$:

$$L = -\log(ML(Q)|_A)$$
$$L_{\text{fan out}} = -\log(ML_{\text{fan out}}(Q)|_{H_{\text{fan out}}(Q)})$$
$$L_{\text{fan in}} = -\log(ML_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))|_{H_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))})$$
$$L_{Adv} = \left| S_1 - \log(ML(SQ_1)|_{A_1}) \right|^2$$
$$L_{Adv_{\text{fan out}}} = |Adv_{\text{fan out}}(Q, (SQ_1, SQ_2)) + L_{\text{fan out}}|^2$$
$$L_{Adv_{\text{fan in}}} = |Adv_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2), A) + L_{\text{fan out}}|^2$$

Note that some of these losses are actually for the same model since $ML_{\text{fan out}}, ML_{\text{fan in}}$ are implemented internally by calling ML and $Adv_{\text{fan out}}, Adv_{\text{fan in}}$ are implemented internally by calling Adv.

Distillation

For the distillation procedure, we want to improve ML by training it to approximate its own ML-consulting-ML tree. Specifically:

1. Generate some $Q \in \mathcal{Q}$ as in amplification.

2. Call $ML_{\text{fan out}}(Q)$. If we get an answer, let $A = ML_{\text{fan out}}(Q)$ and skip to the end. Otherwise, let $SQ_1, SQ_2 = ML_{\text{fan out}}(Q)$.

3. Let $A_1 = ML(SQ_1)$ and $A_2 = ML(SQ_2)$.

4. Call $ML_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$. If we get an answer, let $A = ML_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$. Otherwise, let $SQ_1, SQ_2 = ML_{\text{fan in}}(Q, (Q_1, A_1), (Q_2, A_2))$ and return to step 3.

Then, compute the following losses for ML, Adv:

$$L = -\log(ML(Q)|_A)$$

$$L_{\text{Adv}} = |\text{Adv}(Q, A) + L|^2$$

Iteration

Finally, we want to iterate the above procedure by replacing $H, H_{\text{fan out}}, H_{\text{fan in}}$ in the amplification procedure with some $H', H'_{\text{fan out}}, H'_{\text{fan in}}$. First, let α be some confidence threshold. Then, we will define the following primed Hs:

- Let $H'(Q) = \text{if } \text{Adv}(Q, ML(Q)) + \epsilon < \alpha + \epsilon \text{ then } H(Q) \text{ else } ML(Q)$.
- Let

$$\begin{aligned} & H'_{\text{fan out}}(Q) = \\ & \text{if } \text{Adv}_{\text{fan out}}(Q, ML_{\text{fan out}}(Q)) + \epsilon < \alpha + \epsilon \\ & \ \text{then } H_{\text{fan out}}(Q) \\ & \ \text{else } ML_{\text{fan out}}(Q) \end{aligned}$$

- Let

$$\begin{aligned} & H'_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2)) = \\ & \text{if } \text{Adv}_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2), ML_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2))) + \epsilon < \alpha + \epsilon \\ & \ \text{then } H_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2)) \\ & \ \text{else } ML_{\text{fan in}}(Q, (SQ_1, A_1), (SQ_2, A_2)) \end{aligned}$$

This procedure allows us to continue amplifying the model while using the adversary to require only minimal human data that is selected so as to be maximally helpful.

Conclusion

This proposal differs in a couple of ways from previous proposals made by Paul. First, Paul has recently moved away from discrete amplification/distillation steps. This proposal, however, provides a way to recover discrete steps while still collapsing the recursion. In practice, however, you might still just want to stick with the amplification procedure described here without doing the distillation step, as it isn't strictly necessary.

Second, this proposal uses an adversary to guide the training process. This technique is similar to the concept of importance sampling. The main benefit of this approach is that it takes advantage of active learning by allowing the system to choose which questions and subquestions would be most useful for it to have answered by a human.

Another benefit of the adversary, however, is that it could make transparency much easier. One of the major benefits of IDA is that ML gets trained to approximate its own ML-consulting-ML tree. As a result, understanding the reasoning that went into the final answer produced by ML can be recovered by unfolding its tree (at least in the limit of perfect training). However, unfolding the entire tree is very expensive, as it's linear in the size of the tree. With an adversary, however, you can choose which portions of the tree to unfold first by calling the adversary, enabling you to find errors much more quickly; for a perfect adversary, this reduces the problem of finding an error to $O(\log n)$ instead of $O(n)$.

Thus, the hope is that the use of such an adversary could assist both in making IDA more competitive (by increasing sample efficiency and using active learning) and in making IDA safer (due to the increased ease of transparency).

It should be noted, however, that it is also possible that the use of such an adversary might make the safety situation for IDA worse. First, it introduces the possibility of a robustness to relative scale failure if either ML or Adv gets significantly stronger than the other. One possible way to resolve such an issue, however, might be to give Adv the ability to call ML and vice versa, allowing them to use each other to boost their own capabilities. Second, for an ML and Adv system that are themselves optimizers, with goals that don't perfectly match up with their loss functions, they could cooperate to make it arbitrarily unlikely that H is ever consulted on some specific question. Third, even if ML and Adv weren't cooperating, an RSA-2048-style failure could still prevent the identification of malicious cognition. Resolving failures of these second two types is still an open question (EDIT: see "Risks from Learned Optimization in Advanced Machine Learning Systems," by Hubinger, van Merwijk, Mikulik, Skalse, and Garrabrant).

Can Bayes theorem represent infinite confusion?

Edit: the title was misleading, i didn't ask about a rational agent, but about what comes out of certain inputs in Bayes theorem, so now it's been changed to reflect that.

Eliezer and others talked about how a Bayesian with a 100% prior cannot change their confidence level, whatever evidence they encounter. that's because it's like having infinite certainty. I am not sure if they meant it literary or not (is it really mathematically equal to infinity?), but assumed they do.

I asked myself, well, what if they get evidence that was somehow assigned 100%, wouldn't that be enough to get them to change their mind? In other words -

If P(H) = 100%

And P(E|H) = 0%

than what's P(H|E) equals to?

I thought, well, if both are infinities, what happens when you subtract infinities? the internet answered that it's **indeterminate***, meaning (from what i understand), that it can be anything, and you have absolutely no way to know what exactly.

So i concluded that if i understood everything correct, then such a situation would leave the Bayesian **infinitely confused.** in a state that he has no idea where he is from 0% to a 100%, and no amount of evidence in any direction can ground him anywhere.

Am i right? or have i missed something entirely?

---

*I also found out about Riemann's rearrangement theorem which, in a way, let's you arrange some infinite series in a way that equals whatever you want. Dem, that's cool!

How can we respond to info-cascades? [Info-cascade series]

This is a question in the info-cascade question series. There is a prize pool of up to $800 for answers to these questions. See the link above for full background on the problem (including a bibliography) as well as examples of responses we'd be especially excited to see.

___

In my (Jacob's) work at Metaculus AI, I'm trying to build a centralised space for both finding forecasts as well as the reasoning underlying those forecasts. Having such a space might serve as a simple way for the AI community to avoid runway info-cascades.

However, we are also concerned with situations where new forecasters overweight the current crowd opinion in their forecasts, compared to the underlying evidence, and see this as major risk for the trustworthiness of forecasts to those working in AI safety and policy.

With this question, I am interested in previous attempts to tackle this problem, and how successful they have been. In particular:

- What existing infrastructure has been historically effective for avoiding info-cascades in communities? (Examples could include short-selling to prevent bubbles in asset markets, or norms to share the causes rather than outputs of one's beliefs)
- What problems are not adequately addressed by such infrastructure?

Distribution of info-cascades across fields? [Info-cascade series]

This is a question in the info-cascade question series. There is a prize pool of up to $800 for answers to these questions. See the link above for full background on the problem (including a bibliography) as well as examples of responses we'd be especially excited to see.

___

How common, and how large, are info-cascades in communities that seek to make intellectual progress, such as academia? This distribution is presumably very heavy-tailed as we are dealing with network phenomena. But what is its actual values? How can we estimate this number?

A good starting point for thinking about this might be the paper "How citation distortions create unfounded authority: analysis of a citation network" (Greenberg, 2009), which uses social network theory and graph theory to trace how an at best very uncertain claim in biomedicine cascading into established knowledge. We won't attempt to summarise the paper here.

What societies have ever had legal or accepted blackmail?

Robin Hanson's post History of Blackmail only mentions cases where blackmail was illicit or illegal. Have there ever been any societies, large or small, where blackmail was widely accepted?

AI Safety Prerequisites Course: Basic abstract representations of computation

Followup to AI Safety Prerequisites Course: Revamp and New Lessons. First post.

These are three new lessons of our online course on math formalizations required for AI safety research:

- Level 10: Recursive Functions
- Level 11: Set Theoretic Recursion
- Level 12: The Equivalence of Different Notions of Computability

With these lessons, the student now should:

- Understand the basic abstract representations of computation.
- Know some of what we can expect from computers and also what we can’t expect from them.
- Know a lot more set theoretic tools like equivalence relations and orderings.
- Have seen the construction of the natural numbers from the perspective of set theory.
- Know about mathematical induction, and have used it!
- Know about recursion, and have used it!

If you study using our course, please give us feedback. Leave a comment here or email us at raise@aisafety.info, or through the contact form. Do you have an idea about what prerequisites are most important for AI Safety research? Do you know an optimal way to learn them? Tell us using the same methods or collaborate with us.

Can you check if a mathematical proof is correct? Do you know how to make proofs understandable and easy to remember? Would you like to help to create the prerequisites course? If yes, consider volunteering.

In My Culture

This is a linkpost for https://medium.com/@ThingMaker/in-my-culture-29c6464072b2

Crosspost from Medium; relevant to LessWrong in general and possibly to specific ongoing cultural tensions of the past four months or so. Proposes a simple tool for improving culture-clash dynamics and offers some specifics about the cultural diff between the author and other people. 30min read.

Alignment Newsletter #48

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

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

Highlights

Quantilizers: A Safer Alternative to Maximizers for Limited Optimization and When to use quantilization *(Jessica Taylor and Ryan Carey)*: A key worry with AI alignment is that if we maximize expected utility for some utility function chosen by hand, we will likely get unintended side effects that score highly by the utility function but are nevertheless not what we intended. We might hope to leverage human feedback to solve this: in particular, an AI system that simply mimics human actions would often be desirable. However, mimicry can only achieve human performance, and cannot improve upon it. The first link is a 2015 paper that introduces quantilization, which interpolates between these two extremes to improve upon human performance while bounding the potential (expected) loss from unintended side effects.

In particular, let's suppose that humans have some policy γ (i.e. probability distribution over actions). We evaluate utility or performance using a utility function U, but we *do not assume* it is well-specified -- U can be any function, including one we would not want to maximize. Our goal is to design a policy π that gets higher expected U than γ (reflecting our hope that U measures utility well) without doing too much worse than γ in the worst case when U was as badly designed as possible. We'll consider a one-shot case: π is used to select an action once, and then the game is over.

The core idea behind quantilization is simple: if our policy only does things that the human might have done, any expected loss it incurs corresponds to some loss that the human could incur. So, let's take our human policy γ, keep only the top q-fraction of γ (as evaluated by U), and then sample an action from there. This defines our policy π_q, also called a q-quantilizer. For example, suppose the human would choose A with probability 0.25, B with probability 0.5, and C with probability 0.25, and U(A) > U(B) > U(C). Then a (1/4)-quantilizer would choose A with certainty, a (1/2)-quantilizer would choose randomly between A and B, and a (3/8)-quantilizer would choose A twice as often as B.

Note that even if π_q places all of its probability on the worst possible actions (because U was badly designed), by construction the human had probability q of doing the same thing as π_q, and so the expected cost of π_q can be at most (1/q) times larger than the expected cost of the human policy γ. In fact, if we have no other knowledge, a q-quantilizer maximizes expected utility U subject to the constraint of never doing worse than (1/q) times as bad as γ.

As soon as you move to the setting with multiple actions, if you choose each action using quantilization, then your worst case bound is exponential in the number of actions. If you assume the cost for each action is independent, you recover the guarantees, but this is not a realistic assumption (as the authors note). Long-term plans are very good or very bad *because* all of the actions build on each other to achieve some goal, so the costs are not going to be independent.

The second link is a recent post that further analyzes quantilization. It points out that the theoretical analysis in the paper assumes that deviating from the human policy can only lead to costs. If the deviation could also forgo benefits, then the theorems no longer apply. Perhaps the bottom (1-q) of actions actually have strong positive benefits, that we failed to model in U. Then by eliminating those actions, we may have lost arbitrary amounts of expected value. In these situations, the only way to bound the expected regret is by exact mimicry. It also points out that if you are aiming to simultaneously do well both on U and the worst-case bound, then typically imitation will be better since adding any optimization can drastically weaken the worst-case bound and usually will not make U better by the same amount. Quantilization makes sense when there is a "sweet-spot of actions that are pretty common but substantially outperform imitation".

Rohin's opinion: The exponential blowup in potential loss with multiple actions would make this prohibitive, but of course you could instead view the full sequence of actions (i.e. trajectory) as a mega-action, and quantilize over this mega-action. In this case, a one-millionth-quantilizer could choose from among the million best plans that a human would make (assuming a well-specified U), and any unintended consequences (that were intentionally chosen by the quantilizer) would have to be ones that a human had a one-in-a-million chance of causing to occur, which quite plausibly excludes really bad outcomes.

Phrased this way, quantilization feels like an amplification of a human policy. Unlike the amplification in iterated amplification, it does *not* try to preserve alignment, it simply tries to bound how far away from alignment the resulting policy can diverge. As a result, you can't iterate quantilization to get arbitrarily good capabilities. You might hope that humans could learn from powerful AI systems, grow more capable themselves (while remaining as safe as they were before), and then the next quantilizers would be more powerful.

It's worth noting that the theorem in the paper shows that, to the extent that you think quantilization is insufficient for AI alignment, you need to make some other assumption, or find some other source of information, in order to do better, since quantilization is optimal for its particular setup. For example, you could try to assume that U is at least somewhat reasonable and not pathologically bad; or you could assume an interactive setting where the human can notice and correct for any issues with the U-maximizing plan before it is executed; or you could not have U at all and exceed human performance through some other technique.

I'm not very worried about the issue that quantilization could forgo benefits that the human policy had. It seems that even if this happens, we could notice this, turn off the quantilizer, and fix the utility function U so that it no longer ignores those benefits. (We wouldn't be able to prevent the quantilizer from forgoing benefits of our policy that we didn't know about, but that seems okay to me.)

Technical AI alignment

Iterated amplification

Can HCH epistemically dominate Ramanujan? *(Alex Zhu)*: Iterated amplification rests on the hope that we can achieve arbitrarily high capabilities with (potentially very large) trees of explicit verbal breakdowns of problems. This is often formalized as a question about HCH (AN #34). This post considers the example of Srinivasa Ramanujan, who is "famously known for solving math problems with sudden and inexplicable flashes of insight". It is not clear how HCH would be able to replicate this sort of reasoning.

Learning human intent

Unsupervised Visuomotor Control through Distributional Planning Networks *(Tianhe Yu et al)*

Syntax vs semantics: alarm better example than thermostat *(Stuart Armstrong)*: This post gives a new example that more clearly illustrates the points made in a previous post (AN #26).

Prerequisities: Bridging syntax and semantics, empirically

Interpretability

Synthesizing the preferred inputs for neurons in neural networks via deep generator networks *(Anh Nguyen et al)*

Adversarial examples

Quantifying Perceptual Distortion of Adversarial Examples *(Matt Jordan et al)* (summarized by Dan H): This paper takes a step toward more general adversarial threat models by combining adversarial additive perturbations small in an l_p sense with spatially transformed adversarial examples, among other other attacks. In this more general setting, they measure the size of perturbations by computing the SSIM between clean and perturbed samples, which has limitations but is on the whole better than the l_2 distance. This work shows, along with other concurrent works, that perturbation robustness under some threat models does not yield robustness under other threat models. Therefore the view that l_p perturbation robustness must be achieved before considering other threat models is made more questionable. The paper also contributes a large code library for testing adversarial perturbation robustness.

On the Sensitivity of Adversarial Robustness to Input Data Distributions *(Gavin Weiguang Ding et al)*

Forecasting

Primates vs birds: Is one brain architecture better than the other? *(Tegan McCaslin)*: Progress in AI can be driven by both larger models as well as architectural improvements (given sufficient data and compute), but which of these is more important? One source of evidence comes from animals: different species that are closely related will have similar neural architectures, but potentially quite different brain sizes. This post compares intelligence across birds and primates: while primates (and mammals more generally) have a neocortex (often used to explain human intelligence), birds have a different, independently-evolved type of cortex. Using a survey over non-expert participants about how intelligent different bird and primate behavior is, it finds that there is not much difference in intelligence ratings between birds and primates, but that species with larger brains are rated as more intelligent than those with smaller brains. This only suggests that there are at least two neural architectures that work -- it could still be a hard problem to find them in the vast space of possible architectures. Still, it is some evidence that at least in the case of evolution, you get more intelligence through more neurons, and architectural improvements are relatively less important.

Rohin's opinion: Upon reading the experimental setup I didn't really know which way the answer was going to turn out, so I'm quite happy about now having another data point with which to understand learning dynamics. Of course, it's not clear how data about evolution will generalize to AI systems. For example, architectural improvements probably require some hard-to-find insight which make them hard to find via random search (imagine how hard it would be to invent CNNs by randomly trying stuff), while scaling up model size is easy, and so we might expect AI researchers to be differentially better at finding architectural improvements relative to scaling up model size (as compared to evolution).

Read more: Investigation into the relationship between neuron count and intelligence across differing cortical architectures

Miscellaneous (Alignment)

Quantilizers: A Safer Alternative to Maximizers for Limited Optimization **and** **When to use quantilization** *(Jessica Taylor and Ryan Carey)*: Summarized in the highlights!

Human-Centered Artificial Intelligence and Machine Learning *(Mark O. Riedl)*

AI strategy and policy

Stable Agreements in Turbulent Times *(Cullen O'Keefe)*: On the one hand we would like actors to be able to cooperate before the development of AGI by entering into binding agreements, but on the other hand such agreements are often unpalatable and hard to write because there is a lot of uncertainty, indeterminacy and unfamiliarity with the consequences of developing powerful AI systems. This makes it very hard to be confident that any given agreement is actually net positive for a given actor. The key point of this report is that we can strike a balance between these two extremes by agreeing pre-AGI to be bound by decisions that are made post-AGI with the benefit of increased knowledge. It examines five tools for this purpose: options, impossibility doctrines, contractual standards, renegotiation, and third-party resolution.

Advice to UN High-level Panel on Digital Cooperation *(Luke Kemp et al)*

Other progress in AI

Reinforcement learning

Neural MMO *(OpenAI)* (summarized by Richard): Neural MMO is "a massively multiagent game environment for reinforcement learning agents". It was designed to be persistent (with concurrent learning and no environment resets), large-scale, efficient and expandable. Agents need to traverse an environment to obtain food and water in order to survive for longer (the metric for which they are rewarded), and are also able to engage in combat with other agents. Agents trained within a larger population explore more and consistently outperform those trained in smaller populations (when evaluated together). The authors note that multiagent training is a curriculum magnifier, not a curriculum in itself, and that the environment must facilitate adaptive pressures by allowing a sufficient range of interactions.

Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research *(Joel Z. Leibo, Edward Hughes, Marc Lanctot, Thore Graepel)* (summarized by Richard): The authors argue that the best solution to the problem of task generation is creating multi-agent systems where each agent must adapt to the others. These agents do so first by learning how to implement a high-level strategy, and then by adapting it based on the strategies of others. (The authors use the term "adaptive unit" rather than "agent" to emphasise that change can occur at many different hierarchical levels, and either by evolution or learning). This adaptation may be exogenous (driven by the need to respond to a changing environment) or endogenous (driven by a unit's need to improve its own functionality). An example of the latter is a society implementing institutions which enforce cooperation between individuals. Since individuals will try to exploit these institutions, the process of gradually robustifying them can be considered an automatically-generated curriculum (aka autocurriuclum).

Richard's opinion: My guess is that multiagent learning will become very popular fairly soon. In addition to this paper and the Neural MMO paper, it was also a key part of the AlphaStar training process. The implications of this research direction for safety are still unclear, and it seems valuable to explore them further. One which comes to mind: the sort of deceptive behaviour required for treacherous turns seems more likely to emerge from multiagent training than from single-agent training.

Long-Range Robotic Navigation via Automated Reinforcement Learning *(Aleksandra Faust and Anthony Francis)*: How can we get robots that successfully navigate in the real world? One approach is to use a high-level route planner that uses a learned control policy over very short distances (10-15 meters). The control policy is learned using deep reinforcement learning, where the network architecture and reward shaping is also learned via neural architecture search (or at least something very similar). The simulations have enough noise that the learned control policy transfers well to new environments. Given this policy as well as a floorplan of the environment we want the robot to navigate in, we can build a graph of points on the floorplan, where there is an edge between two points if the robot can safely navigate between the two points using the learned controller (which I *think* is checked in simulation). At execution time, we can find a path to the goal in this graph, and move along the edges using the learned policy. They were able to build a graph for the four buildings at the Google main campus using 300 workers over 4 days. They find that the robots are very robust in the real world. See also Import AI.

Rohin's opinion: This is a great example of a pattern that seems quite common: once we automate tasks using end-to-end training that previously required more structured approaches, new more complex tasks will arise that will use the end-to-end trained systems as building blocks in a bigger structured approach. In this case, we can now train robots to navigate over short distances using end-to-end training, and this has been used in a structured approach involving graphs and waypoints to create robots that can traverse larger distances.

It's also an example of what you can do when you have a ton of compute: for the learned controller, they learned both the network architecture and the reward shaping. About the only thing that had to be explicity specified was the sparse true reward. (Although I'm sure in practice it took a lot of effort to get everything to actually work.)

Competitive Experience Replay *(Hao Liu et al)*

News

Q&A with Jason Matheny, Founding Director of CSET *(Jason Matheny)*: The Center for Security and Emerging Technology has been announced, with a $55 million grant from the Open Philanthropy Project, and is hiring. While the center will work on emerging technologies generally, it will initially focus on AI, since demand for AI policy analysis has far outpaced supply.

One area of focus is the implications of AI on national and international security. Current AI systems are brittle and can easily be fooled, implying several safety and security challenges. What are these challenges, and how important are they? How can we make systems that are more robust and mitigate these problems?

Another area is how to enable effective competition on AI in a global environment, while also cooperating on issues of safety, security and ethics? This will likely require measurement of investment flows, publications, data and hardware across countries, as well as management of talent and knowledge workflows.

See also Import AI.

Rohin's opinion: It's great to see a center for AI policy that's run by a person who has wanted to consume AI policy analysis in the past (Jason Matheny was previously the director of IARPA). It's interesting to see the areas he focuses on in this Q&A -- it's not what I would have expected given my very little knowledge of AI policy.

Show LW: Fallacyfiles

http://www.fallacyfiles.org

What are the fallacy files?

I began collecting and studying logical fallacies about thirty-eight years ago, when I first became interested in logic. This collection took two forms:

A collection of named fallacies—such as "ad hominem"—that is, types of bad reasoning which someone has thought distinctive and interesting enough to name and describe. This collection took the form, primarily, of the study and acquisition of books and articles on the named fallacies, especially textbooks and reference books. You can find individual files on the named fallacies via the Taxonomy of Logical Fallacies, or from the alphabetical index in the scroll bar to your left.

A collection of fallacious, or otherwise bad, arguments, that is, examples of reasoning which may commit one or more of the named fallacies under 1, or are bad in some way yet to be classified. This collection took the form of clippings from newspapers, magazines, pamphlets, photocopies of pages of books, and—in a few rare cases—entire articles or books which were rich sources of bad reasoning. I have used selections from my collection as examples in many of the files on named fallacies, and additional examples can be found in the file: Stalking the Wild Fallacy.

Some years after I began to amass these files, I wondered just what I ought eventually to do with them, how best to organize the information within them, and in what form to make them available to others interested in fallacy studies. The present hypertext web version, The Fallacy Files, was first published on March 11th, 2001, and is the result of this score of years of research and fieldwork on the fallacies.

Willing to share some words that changed your beliefs/behavior?

I'm collecting data on powerfully persuasive speech acts; it's part of a dangling thread of curiosity after GPT-2 (a new and fairly powerful text generation algorithm). I'm skeptical of the danger of mind-warping sentences as sometimes presented in fiction, or AI scenarios, and trying to get a sense of what the territory is like.

I've made a form to collect personal examples of things-someone-said that caused you to seriously change some belief or behavior. An easy example would be if someone declared that they love you, and this caused you to suddenly devote a lot more (or a lot less!) time and attention to them as a person.

If you have five minutes, my goal for this form is 1000+ responses and your own response(s) will help with that. All replies are anonymous, and there's a place for you to restrict how the information is used/state confidentiality desires. You can also fill it out more than once if you want.

https://goo.gl/forms/39x3vJqNomAome382 is the link to the form, if you want to share with anyone else; I'm happy to have this spread around wherever.

[Question] Tracking accuracy of personal forecasts

I've been thinking how I can improve my accuracy predicting events of personal interest (e.g., "Will my landlord get the washing machine fixed within the next two weeks", or "Will my parent die this year" for a more extreme example). Betting markets will not help me with that.

At first I thought about creating dedicated software that gathers such predictions, the final outcomes of predicted events, and presents their accuracy so that the user can spot bias. Then I realised a simple spreadsheet might suffice to gather data at first and assess how useful this is. And if the need arises in the future, it should be easy to import into dedicated software, provided that all the relevant data is already there.

Does anyone track their personal predictions? If so, what methodology do you use, and did it allow you to improve your accuracy?

As an RFC, here's the spreadsheet layout I have on mind:

- Tags: (value 0 or 1):
    - Health
    - Finance
    - Interpersonal relations
    - ...
- Date of the forecast
- Event (e.g., "My landlord will get the washing machine fixed within the next two weeks"). I'm planning to formulate them so that "yes" is always the desired outcome, so that it's easy to spot if I'm reliably too optimistic or pessimistic.
- Estimated probability
- Deadline of the forecast
- Outcome (value 0 or 1, filled after the deadline of the forecast, or when the answer is known sooner)

Boeing 737 MAX MCAS as an agent corrigibility failure

The Boeing Maneuvering Characteristics Augmentation System (MCAS) can be thought of, if reaching a bit, as a specialized AI: it performs a function normally reserved for a human pilot: pitching the nose down when it deems the angle of attack to be dangerously high. This is not, by itself, a problem. There are pilots in the cockpit who can take control when needed.

Only in this case they couldn't. Simply manually pitching the nose up when the MCAS pitching it down too much would not disengage the system, it would activate again, and again. One has to manually disengage the autopilot (this information was not in the pilot training). For comparison, think of the cruise control system in a car: the moment you press the brake pedal, it disengages; if you push the gas pedal, then release, it return to the preset speed. At no time it tries to override your actions. Unlike MCAS.

MCAS disregards critical human input and even fights the human for control in order to reach its goal of "nominal flight parameters". From the Corrigibility paper:

> We say that an agent is “corrigible” if it tolerates or assists many forms of outside correction, including atleast the following: (1) A corrigible reasoner must at least tolerate and preferably assist the programmers in their attempts to alter or turn off the system...

In this case the "agent" actively fought its human handlers instead of assisting them. Granted, the definition above is about programmers, not pilots, and the existing MCAS probably would not fight a software update, being a dumb specialized agent.

But we are not that far off: a lot of systems include built-in security checks for the remote updates. If one of those checks were to examine the algorithm the updated code uses and reject it when it deems it unacceptable because it fails its internal checks, the corrigibility failure would be complete! In a life-critical always-on system this would produce a mini-Skynet. I don't know whether something like that has happened yet, but I would not be surprised if it has, and resulted in catastrophic consequences.

Comparison of decision theories (with a focus on logical-counterfactual decision theories)

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

Introduction

Summary

This post is a comparison of various existing decision theories, with a focus on decision theories that use logical counterfactuals (a.k.a. the kind of decision theories most discussed on LessWrong). The post compares the decision theories along outermost iteration (action vs policy vs algorithm), updatelessness (updateless or updateful), and type of counterfactual used (causal, conditional, logical). It then explains the decision theories in more detail, in particular giving an expected utility formula for each. The post then gives examples of specific existing decision problems where the decision theories give different answers.

Value-added

There are some other comparisons of decision theories (see the “Other comparisons” section), but they either (1) don’t focus on logical-counterfactual decision theories; or (2) are outdated (written before the new functional/logical decision theory terminology came about).

To give a more personal motivation, after reading through a bunch of papers and posts about these decision theories, and feeling like I understood the basic ideas, I remained highly confused about basic things like “How is UDT different from FDT?”, “Why was TDT deprecated?”, and “If TDT performs worse than FDT, then what’s one decision problem where they give different outputs?” This post hopes to clarify these and other questions.

None of the decision theory material in this post is novel. I am still learning the basics myself, and I would appreciate any corrections (even about subtle/nitpicky stuff).

Audience

This post is intended for people who are similarly confused about the differences between TDT, UDT, FDT, and LDT. In terms of reader background assumed, it would be good to know the statements to some standard decision theory problems (Newcomb’s problem, smoking lesion, Parfit’s hitchhiker, transparent box Newcomb’s problem, counterfactual mugging (a.k.a. curious benefactor; see page 56, footnote 89)) and the "correct" answers to them, and having enough background in math to understand the expected utility formulas.

If you don't have the background, I would recommend reading chapters 5 and 6 of Gary Drescher's *Good and Real* (explains well the idea of subjunctive means–end relations), the FDT paper (explains well how FDT's action selection variant works, and how FDT differs from CDT and EDT), "Cheating Death in Damascus", and "Toward Idealized Decision Theory" (explains the difference between policy selection and logical counterfactuals well), and understanding what Wei Dai calls "decision theoretic thinking" (see comments: 1, 2, 3). I think a lot of (especially old) content on decision theory is confusingly written or unfriendly to beginners, and would recommend skipping around to find explanations that "click".

Comparison dimensions

My main motivation is to try to distinguish between TDT, UDT, and FDT, so I focus on three dimensions for comparison that I think best display the differences between these decision theories.

Outermost iteration

All of the decision theories in this post iterate through some set of "options" (intentionally vague) at the outermost layer of execution to find the best "option". However, the nature (type) of these "options" differs among the various theories. Most decision theories iterate through either *actions* or *policies*. When a decision theory iterates through actions (to find the best action), it is doing "action selection", and the decision theory outputs a single action. When a decision theory iterates through policies (to find the best policy), it is doing "policy selection", and outputs a single *policy*, which is an observation-to-action mapping. To get an action out of a decision theory that does policy selection (because what we really care about is knowing which action to take), we must *call* the policy on the actual observation.

Using the notation of the FDT paper, an action has type $\mathcal{A}$ while a policy has type $\mathcal{X} \to \mathcal{A}$, where $\mathcal{X}$ is the set of observations. So given a policy $\pi : \mathcal{X} \to \mathcal{A}$ and observation $x \in \mathcal{X}$, we get the action by calling π on x, i.e. $\pi(x) \in \mathcal{A}$.

From the expected utility formula of the decision theory, you can tell action vs policy selection by seeing what variable comes beneath the $\arg\max$ operator (the $\arg\max$ operator is what does the outermost iteration); if it is $a \in \mathcal{A}$ (or similar) then it is iterating over actions, and if it is $\pi \in \Pi$ (or similar), then it is iterating over policies.

One exception to the above is UDT2, which seems to iterate over *algorithms*.

Updatelessness

In some decision problems, the agent makes an observation, and has the choice of updating on this observation before acting. Two examples of this are: in counterfactual mugging (a.k.a. curious benefactor), where the agent makes the observation that the coin has come up tails; and in the transparent box Newcomb's problem, where the agent sees whether the big box is full or empty.

If the decision algorithm updates on the observation, it is *updateful* (a.k.a. "not updateless"). If it doesn't update on the observation, it is *updateless*.

This idea is similar to how in Rawls's "veil of ignorance", you must pick your moral principles, societal policies, etc., before you find out who you are in the world or as if you don't know who you are in the world.

How can you tell if a decision theory is updateless? In its expected utility formula, if it conditions on the observation, it is updateful. In this case the probability factor looks like $P(\ldots \mid \ldots, \textsc{Obs} = x)$, where x is the observation (sometimes the observation is called "sense data" and is denoted by s). If a decision theory is updateless, the conditioning on "$\textsc{Obs} = x$" is absent. Updatelessness only makes a difference in decision problems that have observations.

There seem to be different meanings of "updateless" in use. In this post I will use the above meaning. (I will try to post a question on LessWrong soon about these different meanings.)

Type of counterfactual

In the course of reasoning about a decision problem, the agent can construct counterfactuals or hypotheticals like "if I do *this*, then *that* happens". There are several different kinds of counterfactuals, and decision theories are divided among them.

The three types of counterfactuals that will concern us are: causal, conditional/evidential, and logical/subjunctive. The distinctions between these are explained clearly in the FDT paper so I recommend reading that (and I won't explain them here).

In the expected utility formula, if the probability factor looks like $P(\ldots \mid \ldots, \textsc{Act} = a)$ then it is evidential, and if it looks like $P(\ldots \mid \ldots, \mathrm{do}(\textsc{Act} = a))$ then it is causal. I have seen the logical counterfactual written in many ways:

- $P(\ldots \mid \ldots, \mathrm{do}(\textsc{dt}(\ldots) = \ldots))$ e.g. in the FDT paper, p. 14
- $P(\ldots \mid \ldots, \mathrm{true}(\textsc{dt}(\ldots) = \ldots))$ e.g. in the FDT paper, p. 14
- $P(\ulcorner \textsc{DT}(\ldots) = \ldots \urcorner \mathrel{\Box\!\!\rightarrow} \ldots \mid \ldots)$ e.g. in Hintze, p. 4
- $P(\ulcorner \textsc{DT}(\ldots) = \ldots \urcorner \triangleright \ldots \mid \ldots)$ e.g. on Arbital

Other dimensions that I ignore

There are many more dimensions along which decision theories differ, but I don't understand these and they seem less relevant for comparing among the main logical-counterfactual decision theories, so I will just list them here but won't go into them much later on in the post:

- Reflective consistency (in particular dynamic consistency): I think this is about whether an agent would use precommitment mechanisms or self-modify to use a different decision theory. Can this be seen immediately from the expected utility formula? If not, it might be unlike the other three above. My current guess is that reflective consistency is a higher-level property that follows from the above three.
- Emphasis on graphical models: FDT is formalized using graphical models (of the kind you can read about in Judea Pearl's book *Causality*) while UDT isn't.
- Recent developments like using logical inductors.
- Uncertainty about where your decision algorithm is: I think this is some combination of the three that I'm already covering. For previous discussions, see this section of Andrew Critch's post, this comment by Wei Dai, and this post by Vladimir Slepnev.
- Different versions of UDT (e.g. proof-based, modal).

Comparison table along the given dimensions

Given the comparison dimensions above, the decision theories can be summarized as follows:

Decision theory	Outermost iteration	Updateless	Type of counterfactual
Updateless decision theory 1 (UDT1)	action	yes	logical
Updateless decision theory 1.1 (UDT1.1)	policy	yes	logical
Updateless decision theory 2 (UDT2)	algorithm	yes	logical
Functional decision theory, iterating over actions (FDT-action)	action	yes	logical
Functional decision theory, iterating over policies (FDT-policy)	policy	yes	logical
Logical decision theory (LDT)	unspecified	unspecified	logical
Timeless decision theory (TDT)	action	no	logical
Causal decision theory (CDT)	action	no	causal
Evidential decision theory (EDT, "naive EDT")	action	no	conditional

The general "shape" of the expected utility formulas will be:

$$\arg\max_{\substack{\text{outermost}\\ \text{iteration}}} \sum_{j=1}^{N} U(o_j) \cdot P(\textsc{Outcome} = o_j \mid \text{updatelessness}, \text{counterfactual})$$

Or sometimes:

$$\arg\max_{\substack{\text{outermost}\\ \text{iteration}}} \sum_{j=1}^{N} U(o_j) \cdot P(\text{counterfactual} \mathrel{\Box\!\!\rightarrow} \textsc{Outcome} = o_j \mid \text{updatelessness})$$

Explanations of each decision theory

This section elaborates on the comparison above by giving an expected value formula for each decision theory and explaining why each cell in the table takes that particular value. I won't define the notation very clearly, since I am mostly collecting the various notations that have been used (so that you can look at the linked sources for the details). My goals are to explain how to fill in the table above and to show how all the existing variants in notation are saying the same thing.

UDT1 and FDT (iterate over actions)

I will describe UDT1 and FDT's action variant together, because I think they give the same decisions (if there's a decision problem where they differ, I would like to know about it). The main differences between the two seem to be:

1. The way they are formalized, where FDT uses graphical models and UDT1 uses some kind of non-graphical "mathematical intuition module".
2. The naming, where UDT1 emphasizes the "updateless" aspect and FDT emphasizes the logical counterfactual aspect.
3. Some additional assumptions that UDT has that FDT doesn't. Rob Bensinger says "accepting FDT doesn't necessarily require a commitment to some of the philosophical ideas associated with updatelessness and logical prior probability that MIRI, Wei Dai, or other FDT proponents happen to accept" and also says UDT "built in some debatable assumptions (over and above what's needed to show why TDT, CDT, and EDT don't work)". I'm not sure what these additional assumptions are, but my guess is it has to do with viewing the world as a program, Tegmark's level IV multiverse, and things like that (I would be interested in hearing more about the exact assumptions).

In the original UDT post, the expected utility formula is written like this:

$$Y^* = \arg\max_{Y} \sum P_Y(\langle E_1, E_2, E_3, \ldots \rangle) U(\langle E_1, E_2, E_3, \ldots \rangle)$$

Here Y is an “output string” (which is basically an action). The sum is taken over all possible vectors of the execution histories. I prefer Tyrrell McAllister’s notation:

$$\arg\max_{Y\in\mathbf{Y}} \sum_{E\in\mathbf{E}} M(X, Y, E)U(E)$$

To explain the UDT1 row in the comparison table, note that:

- The outermost iteration is $\arg\max_{Y\in\mathbf{Y}}$ (over output strings, a.k.a. actions), so it is doing action selection.
- We don’t update on the observation. This isn’t really clear from the notation, since $M(X, Y, E)$ still depends on the input string X. However, the original post clarifies this, saying “Bayesian updating is not done explicitly in this decision theory”.
- The counterfactual is logical because P_Y and M use the “mathematical intuition module”.

In the FDT paper (p. 14), the action selection variant of FDT is written as follows:

$$\begin{aligned}\mathrm{FDT}(P, G, x) &= \arg\max_{a\in\mathcal{A}} \mathbb{E}(\mathcal{U}(\textsc{Outcome}) \mid \mathtt{do}(\textsc{fdt}(\underline{P}, \underline{G}, \underline{x}) = a)) \\ &= \arg\max_{a\in\mathcal{A}} \sum_{j=1}^{N} \mathcal{U}(o_j) \cdot P(\textsc{Outcome} = o_j \mid \mathtt{do}(\textsc{fdt}(\underline{P}, \underline{G}, \underline{x}) = a))\end{aligned}$$

Again, note that we are doing action selection (“$\arg\max_{a\in\mathcal{A}}$”), using logical counterfactuals (“$\mathtt{do}(\textsc{fdt}(\underline{P}, \underline{G}, \underline{x}) = a)$”), and being updateless (absence of “$\textsc{Obs} = x$”).

UDT1.1 and FDT (iterate over policies)

UDT1.1 is a decision theory introduced by Wei Dai’s post “Explicit Optimization of Global Strategy (Fixing a Bug in UDT1)”.

In Hintze (p. 4, 12) UDT1.1 is written as follows:

$$\mathrm{UDT}(s) = \arg\max_{f} \sum_{i=1}^{n} U(O_i) \cdot P(\ulcorner \mathrm{UDT} := f : s \mapsto a \urcorner \;\Box\!\!\rightarrow O_i)$$

Here f iterates over functions that map sense data (s) to actions (a), U is the utility function, and $O_1, \ldots, O_n$ are outcomes.

Using Tyrrell McAllister's notation, UDT1.1 looks like:

$$\mathrm{UDT}_{1.1}(X, Y, E, M, I) = \arg\max_{f \in I} \sum_{E \in E} M(f, E)U(E)$$

Using notation from the FDT paper plus a trick I saw on this Arbital page we can write the policy selection variant of FDT as:

$$(\mathrm{FDT}(P, x))(x) = \left(\arg\max_{\pi \in \Pi} \sum_{j=1}^{N} U(o_j) \cdot P(\textsc{Outcome} = o_j \mid \mathrm{true}(\textsc{fdt}(\underline{P}, \underline{x}) = \pi))\right)(x)$$

On the right hand side, the large expression (the part inside and including the $\arg\max$) returns a policy, so to get the action we call the policy on the observation x.

The important things to note are that UDT1.1 and the policy selection variant of FDT:

- Do policy selection because the outermost iteration is over policies ("$\arg\max_f$" or "$\arg\max_{\pi \in \Pi}$" depending on the notation). Quotes about policy selection: The FDT paper (p. 11, footnote 7) says "In the authors' preferred formalization of FDT, agents actually iterate over *policies* (mappings from observations to actions) rather than actions. This makes a difference in certain multi-agent dilemmas, but will not make a difference in this paper." See also comments by Vladimir Slepnev (1, 2).
- Use logical counterfactuals (denoted by corner quotes and boxed arrow, the mathematical intuition M, or the true operator).
- Are updateless because they don't condition on the observation (note the absence of conditioning of the form $\textsc{Obs} = x$).

TDT

My understanding of TDT is mainly from Hintze. I am aware of the TDT paper and skimmed it a while back, but did not revisit it in the course of writing this post.

Using notation from Hintze (p. 4, 11) the expected utility formula for TDT can be written as follows:

$$\mathrm{TDT}(s) = \arg\max_{a \in A} \sum_{i=1}^{n} U(O_i) P(\ulcorner \mathrm{TDT}(s) := a \urcorner \mathbin{\Box\!\!\rightarrow} O_i \mid s)$$

Here, s is a string of sense data (a.k.a. observation), A is the set of actions, U is the utility function, $O_1, \ldots, O_n$ are outcomes, the corner quotes and boxed arrow $\Box\!\!\rightarrow$ denote a logical counterfactual ("if the TDT algorithm were to output a given input s").

If I were to rewrite the above using notation from the FDT paper, it would look like:

$$\mathrm{TDT}(P, x) = \arg\max_{a \in A} \sum_{j=1}^{N} U(o_j) \cdot P(\textsc{Outcome} = o_j \mid \textsc{Obs} = x, \mathrm{true}(\textsc{tdt}(\underline{P}, \underline{x}) = a))$$

The things to note are:

- The outermost iteration is over actions ("$\arg\max_{a \in A}$"), so TDT does action selection.
- We condition on the sense data s or observation $\textsc{Obs} = x$, so TDT is updateful.

  Quotes about TDT's updatefulness: this post describes TDT as "a theory by MIRI senior researcher Eliezer Yudkowsky that made the mistake of conditioning on observations". The Updateless decision theories page on Arbital calls TDT "updateful". Hintze (p. 11): "TDP's failure on the Curious Benefactor is straightforward. Upon seeing the coinflip has come up tails, it updates on the sensory data and realizes that it is in the causal branch where there is no possibility of getting a million."
- We use corner quotes and the boxed arrow, or the true operator, to denote a logical counterfactual.

UDT2

I know very little about UDT2, but based on this comment by Wei Dai and this post by Vladimir Slepnev, it seems to iterate over algorithms rather than actions or policies, and I am assuming it didn't abandon updatelessness and logical counterfactuals.

The following search queries might have more information:

- "UDT2"
- site:agentfoundations.org "UDT2"
- site:lesswrong.com "UDT2"

LDT

LDT (logical decision theory) seems to be an umbrella decision theory that only requires the use of logical counterfactuals, leaving the iteration type and updatelessness unspecified. So my understanding is that UDT1, UDT1.1, UDT2, FDT, and TDT are all logical decision theories. See this Arbital page, which says:

> "Logical decision theories" are really a family of recently proposed decision theories, none of which stands out as being clearly ahead of the others in all regards, but which are allegedly all better than causal decision theory.

The page also calls TDT a logical decision theory (listed under "non-general but useful logical decision theories").

CDT

Using notation from the FDT paper (p. 13), we can write the expected utility formula for CDT as follows:

$$\begin{aligned}\mathrm{CDT}(P, G, x) &= \operatorname*{arg\,max}_{a \in A} \mathbb{E}(U(\textsc{Outcome}) \mid do(\textsc{Act} = a), \textsc{Obs} = x) \\ &= \operatorname*{arg\,max}_{a \in A} \sum_{j=1}^{N} U(o_j) \cdot P(\textsc{Outcome} = o_j \mid do(\textsc{Act} = a), \textsc{Obs} = x)\end{aligned}$$

Things to note:

- The outermost iteration is $\operatorname{arg\,max}_{a \in A}$ so CDT does action selection.
- We condition on $\textsc{Obs} = x$ so CDT is updateful.
- The presence of $do(\textsc{Act} = a)$ means we use causal counterfactuals.

EDT

Using notation from the FDT paper (p. 12), we can write the expected utility formula for EDT as follows:

$$\begin{aligned}\mathrm{EDT}(P, x) &= \operatorname*{arg\,max}_{a \in A} \mathbb{E}(U(\textsc{Outcome}) \mid \textsc{Obs} = x, \textsc{Act} = a) \\ &= \operatorname*{arg\,max}_{a \in A} \sum_{j=1}^{N} U(o_j) \cdot P(\textsc{Outcome} = o_j \mid \textsc{Obs} = x, \textsc{Act} = a)\end{aligned}$$

Things to note:

- The outermost iteration is $\arg\max_{a \in A}$ so EDT does action selection.
- We condition on $\textsc{Obs} = x$ so EDT is updateful.
- We condition on $\textsc{Act} = a$ so EDT uses conditional probability as its counterfactual.

There are various versions of EDT (e.g. versions that smoke on the smoking lesion problem). The EDT in this post is the “naive” version. I don’t understand the more sophisticated versions of EDT, but the keyword for learning more about them seems to be the tickle defense.

Comparison on specific decision problems

If two decision theories are actually different, there should be some decision problem where they return different answers.

The FDT paper does a great job of distinguishing the logical-counterfactual decision theories from EDT and CDT. However, it doesn’t distinguish between different logical-counterfactual decision theories.

The following is a table that shows the disagreements between decision theories. For each pair of decision theories specified by a row and column, the decision problem named in the cell is one where the decision theories return different answers. The diagonal is blank because the decision theories are the same. The lower left triangle is blank because it repeats the entries in the mirror image (along the diagonal) spots.

	UDT1.1/FDT-policy	**UDT1/FDT-action**	**TDT**	**EDT**	**CDT**
UDT1.1/FDT-policy	–	Number assignment problem described in the UDT1.1 post (both UDT1 copies output “A”, the UDT1.1 copies output “A” and “B”)	Counterfactual mugging (a.k.a. curious benefactor) (TDT refuses, UDT1.1 pays)	Parfit’s hitchhiker (EDT refuses, UDT1.1 pays)	Newcomb’s problem (CDT two-boxes, UDT1.1 one-boxes)
UDT1/FDT-action	–	–	Counterfactual mugging (a.k.a. curious benefactor) (TDT refuses, UDT1 pays)	Parfit’s hitchhiker (EDT refuses, UDT1 pays)	Newcomb’s problem (CDT two-boxes, UDT1 one-boxes)

	UDT1.1/FDT-policy	**UDT1/FDT-action**	**TDT**	**EDT**	**CDT**
TDT	–	–	–	Parfit’s hitchhiker (EDT refuses, TDT pays)	Newcomb’s problem (CDT two-boxes, TDT one-boxes)
EDT	–	–	–	–	Newcomb’s problem (CDT two-boxes, EDT one-boxes)
CDT	–	–	–	–	–

Other comparisons

Here are some existing comparisons between decision theories that I found useful, along with reasons why I felt the current post was needed.

- “Decision-theoretic problems and Theories; An (Incomplete) comparative list” by somervta. This list is useful and modern but doesn’t include the different versions of UDT and FDT.
- “A comprehensive list of decision theories” by Caspar Oesterheld and/or Johannes Treutlein. I think my motivation is different from that of the author(s) of this list; I mainly want to distinguish between all the UDTs, TDT, and FDT, so my tables and columns of those tables are chosen in a way so as to make the differences apparent.
- “Problem Class Dominance in Predictive Dilemmas” by Daniel Hintze. This paper is from 2014 so doesn’t include the FDT/LDT terminology, and also doesn’t include the various versions of UDT.
- “Timeline of decision theory”. This is an incomplete timeline I’ve been working on sporadically. It gives a chronological ordering of some decision theories and decision problems with a focus on logical-counterfactual decision theories, but doesn’t really compare them.

Declarative Mathematics

Programmers generally distinguish between “imperative” languages in which you specify *what to do* (e.g. C) versus “declarative” languages in which you specify *what you want*, and let the computer figure out how to do it (e.g. SQL). Over time, we generally expect programming to become more declarative, as more of the details are left to the compiler/interpreter. Good examples include the transition to automated memory management and, more recently, high-level tools for concurrent/parallel programming.

It’s hard to say what programming languages will look like in twenty or fifty years, but it’s a pretty safe bet that they’ll be a lot more declarative.

I expect that applied mathematics will also become much more declarative, for largely the same reasons: as computers grow in power and software expands its reach, there will be less and less need for (most) humans to worry about the details of rote computation.

What does this look like? Well, let’s start with a few examples of “imperative” mathematics:

- Most grade-school arithmetic: it’s explicitly focused on computation, and even spells out the exact steps to follow (e.g. long division).
- Gaussian reduction, as typically taught in a first-semester linear algebra class. It’s the undergrads’ version of grade-school arithmetic.
- Most of the computation performed by hand in physics, engineering and upper-level econ courses & research, i.e. algebra/DEs/PDEs.

Contrast to the declarative counterparts:

- Figure out what arithmetic needs to be done (i.e. what numbers to plug in) and then use a calculator
- Set up a system of linear equations, then have python or wolfram invert the matrix
- Choose which phenomena to include in a model, set up the governing equations, then use either numerical simulation (for pretty graphs) or a computer algebra system (for asymptotics and scaling relations).

In the declarative case, most of the work is in formulating the problem, figuring out what questions to ask, and translating it all into a language which a computer can work with - numbers, or matrices, or systems of equations.

This is all pretty standard commentary at the level of mathematics education, but the real importance is in shaping the *goals* of applied mathematics. For the past century, the main objectives of mathematical research programs would be things like existence & uniqueness, or exhaustive classification of some objects, or algorithms for solving some problem (a.k.a. constructive solution/proof). With the shift toward declarative mathematics, there will be more focus on *building declarative frameworks* for solving various kinds of problems.

The best example I know of is convex analysis, in the style taught by Stephen Boyd (course, book). Boyd’s presentation is the user’s guide to convex optimization: it addresses what kinds of questions can be asked/answered, how to recognize relevant applications in the wild, how to formulate problems, what guarantees are offered in terms of solutions, and of course a firehose of examples from a wide variety of fields. In short, it includes exactly the pieces needed to use the tools of convex analysis as a declarative framework. By contrast, the internals of optimization algorithms are examined only briefly, with little depth and a focus on things which a user might need to tweak. Complicated proofs are generally omitted altogether, the relevant results simply stated as tools available for use.

This is what a mature declarative mathematical framework looks like: it provides a set of tools for practitioners to employ on practical problems. Users don't need to know what's going on under the hood; the algorithms and proofs generally "just work" without the user needing to worry about the details. The user's job is to understand the language of the framework, the interface, and translate their own problems into that language. Once they've expressed what they want, the tools take over and handle the rest.

That's the big goal of future mathematical disciplines: provide a practical framework which practitioners can use to solve real-world problems in the wild, without having to know all the little details and gotchas under the hood.

One last example, which is particularly relevant to me and to ML/AI research. One of the overarching goals of probability/statistics/ML is to be able to code up a generative model, pass it into a magical algorithm, and get back parameter estimates and uncertainties. The "language" of generative models is very intuitive and generally easy to work with, making it an excellent interface to a declarative mathematical toolkit. Unfortunately, the behind-the-scenes part of the toolkit remains relatively finicky and inefficient. As of today, the "magical algorithm" part is usually MCMC, which is great in terms of universality but often super-exponentially slow for multimodal problems, especially in high dimensions, and can converge very slowly even in simple unimodal problems. It's not really reliable enough to use without thinking about what's under the hood. Better mathematical tools and guarantees are needed before this particular framework fully matures.

If anyone has other examples of maturing or up-and-coming declarative mathematical frameworks, I'd be very interested to hear about them.

If you wrote a letter to your future self every day, what would you put in it?

Several days ago, I wrote an email to myself.

That email will now be sent to me every day.

All it is is a single draft in my Gmail drafts folder, with the Mail Conductor extension sending it out at 10:00 am. I can modify the draft whenever I want, each time improving it.

Consider, with the fervent munchkinry of a final exam... What would *you* send yourselves?

(Helpful anchor point: What would you share with a guaranteed audience of thousands of cooperative strangers who thought *very much* - but not *quite* totally - like you?)

Insights from Munkres' Topology

This is about the Math Textbook **Topology** from Miri's research guide. (You can find the pdf online for free.) I got this book about a year ago. It takes a rigorous bottom-up approach that requires almost no prior knowledge but a lot of time. It's long and there are many exercises. I've read most of the book and done most of the exercises in the parts I read. It taught me about topology, about proving theorems, and about being efficient with Latex.

Chapter 1: Set Theory and Logic

This is a general introduction to highest mathematics and has nothing to do with topology. It introduces fundamental concepts such as logical implications, sets, tuples, relations, and functions. I've worked through this perhaps more thoroughly than I needed to, but I got some real value out of it: the book makes some things explicit that are often brushed over, such as when and why one is allowed to use proofs by induction, or what hides behind the supremum operation on an ordered set, and when one is allowed to use it.

The most interesting part about this was the construction of the usual number sets. Rather than beginning by defining N (for example through $n = \{0, ..., n-1\}$ as is done in ZFC), it starts by asserting the existence of a set R called the real numbers, and of two operators $+, \cdot : R^2 \to R$ and an order relation $<$ on R which fulfil a list of eight axioms. From there, the sets Z and Q are constructed out of R.

Two of the axioms on R state that $<$ has the least-upper-bound property (which is precisely what is needed for the supremum) and that, given $x < y$ in R, there is an element $z \in R$ such that $x < z < y$.

This approach is quite different from the ZFC construction: now R is taken to be the most fundamental set rather than something one needs to be constructed through a sequence of complicated steps. This intuition is compatible with the rest of the book: as a topological space, R is a more standard example than N. The approach also requires less work.

Chapter 2: Topological Spaces and Continuous Functions

A **topological space** is a pair (X, T) where X is any set and T a set of subsets of X, that is, $T \in P(P(X))$, or equivalently $T \subset P(X)$. One can think of the topology as a bunch of bubbles covering the elements of X. A topology must meet the following three axioms:

(1): $\emptyset, X \in T$

(2): $\forall O_1, ..., O_n \in T : (O_1 \cap ... \cap O_n) \in T$

(3): $\forall \{O_j\}_{j \in J} \subset T : (\bigcup_{j \in J} O_j) \in T$

That is, the topology is closed under *arbitrary* unions (3) (J is any index set), but only *finite* intersections (2).

A subset $O \subset X$ is called **open** if and only if $O \in T$, and it is called **closed** if and only if $(X - O) \subset T$ (the $-$ is set-difference). Open is not the opposite of closed; a set can be open or closed or neither of both (like $\emptyset$ and X). A topology just is then just the collection of all open sets of X. Before reading this book, insofar as I knew what open sets were at all, I used to think of them as sets where every point has a small area around it that is also in the set (such as $\{(x, y) \in R : x^2 + y^2 < 1\}$). But the topological definition also allows T to be $\emptyset$ plus all sets that contain a fixed point $x_0 \in X$, for example. As far as I know, this topology is not seriously "used" for anything, but it does meet all three axioms. Other strange examples exist.

The book mentions that it took a while to reach a consensus on what exactly a topological space should be, and this appears to be the most useful generalization of concepts from analysis. Here is a theorem which I find gives it a bit of intuition.

Theorem. Let (X, T) be a topological space, and let $O \subset X$. Then,

$$O \in T \iff \forall x \in O \ \exists U \in T : x \in U \subset O$$

This reads "A set O is open if and only if for each of its points, there exists an open set around that point which is contained in O". This might be closer to what one thinks being open means.

Proof. ‘‘ $\Longrightarrow$ ": given $x \in O$, one has $x \in O \subset O$. ‘‘ $\Longleftarrow$ ": for each $x \in X$, let U_x be an open set such that $x \in U_x \subset O$. Then $O = \bigcup_{x \in X} U_x$, so O is open. //

The second step is using the fact that arbitrary unions of open sets are open. I remember feeling intuitively that if a bubble is put around every element in the set, then the union of all of these bubbles must be more than just the set itself. But if the bubbles are all contained in the set, then it's easy to prove that the union is precisely the set itself.

This theorem has frequent use: most of the time one wants to show that a set A is open, one does it by picking an arbitrary element in the set and fitting an open set around it while staying within A.

The standard topology on R consists of all the sets that are unions of open intervals (a, c) with $a < c$ in R. Note that the intersection of all open sets containing a point $x \in R$ is just the one-point set $\{x\}$ (this follows from the second axiom of R that I listed). But the intersection of arbitrarily many points does not need to be open; only finite intersections need be open And indeed, $\{x\}$ is not open in R (but it is closed).

One way to define a topology on a space X is to define a metric $d : X^2 \rightarrow R$ that meets a bunch of properties and is supposed to be a coherent measure of distance between any two points of X. The topology induced by d consists of all sets $O \subset X$ for which there is an $\epsilon \in R$ such that $B_d(x, \epsilon) := \{p \in X : d(x, p) < \epsilon\} \subset O$. Then (X, d) is called a *metric space,* and given a function between two metric spaces, one can define continuity with the ϵ-δ definition from analysis. But the topological definition of continuity is more general than that. Given a function $f : X \rightarrow Y$ between two topological spaces, f is defined to be continuous if and only if for every set O open in Y, the set $f^{-1}(O)$ is open in X. This is equivalent to the ϵ-δ definition in cases where X is a metric with the topology induced by d. It's more general because every metric induces a topology, but not every topology has a metric inducing it.

In most fields of mathematics, functions are a central focus (why is this?). Most (all?) fields take a special interest in some particular class of functions, which are usually just a tiny area in the space of all functions (take a continuous function from R to R and change any image point by any amount, and it's no longer continuous). In algebra, one wants to have functions that preserve *structure*; in the of a function between two abstract groups, one wants that $f(x * y) = f(x) * f(y)$. A function fulfilling this is then called a **homomorphism**. The analogous concept in topology and analysis is a **continuous function**: it doesn't preserve structure (there need not be any "structure" analogous to that induced by the operator $*$ on a topological space), but it preserves *topology*, which for metric spaces means that arbitrarily small changes of x lead to arbitrarily small spaces of $f(x)$, and in the more general case of topology, that for every open set U around $f(x)$ there must be an open set O around x such that

$f(O) \subset U$ (this is equivalent to the requirement that $f^{-1}(U)$ be open, by the theorem I proved above). The analogous concept to an **isomorphism** between groups, which is a bijective homomorphism, is then a **homeomorphism**, which is a bijective continuous function f such that f^{-1} is also continuous. If there exists a homeomorphism $f : X \to Y$ between two topological spaces X, and Y, then they are called **homeomorphic**. In that case, they are said to be "topologically identical", since all properties which are formulated in terms of their topologies (such as the existence of continuous functions that do certain things) are equivalent for both. There are many such properties that are of interest.

This has always been somewhat unintuitive to me. Whether two spaces are homeomorphic depends on seemingly strange things; for example, the open interval $(0, 1)$ (the term open has a different meaning for intervals than for sets in a topological space, but if $\mathbb{R}$ is given the standard topology, they coincide) which might intuitively seem "small" is homeomorphic to the entire set of real numbers $\mathbb{R}$. Concepts like length are not topological; they can change under a homeomorphism. But fine, that's similar to familiar properties of infinity: the sets $\mathbb{N}$ and $\mathbb{Q}$ don't feel like they're the same size, but they're bijective. Same for $(0, 1)$ and $\mathbb{R}$. However, the half-open interval [0,1), it is no longer homeomorphic to $\mathbb{R}$. A single point has changed things, which is new: the spaces $[0, 1)$ and $\mathbb{R}$ are still bijective (even though writing down an explicit bijective function is tricky).

In the real world, if we go down to the smallest building blocks of the universe, then their impact also becomes arbitrarily small, I believe. This makes it seem implausible that a formal system where single points have so much importance is useful. But obviously, this intuition is wrong. For example, fixed point theorems seem to be of some importance in AI alignment, and those are fundamentally topological problems. The *disc* (that is, the space $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$) is often denoted D^2, and one can prove that any continuous function $f : D^2 \to D^2$ has a fixed point, that is, there exists a point $x \in D^2$ such that $f(x) = x$. This is a result that falls out of deeper studies of topology (though there are many different ways to prove it), and it also generalizes to the n-dimensional ball D^n. Once again, it is no longer true when one takes a point out of D^n (if one takes out the center, for example, then the function rotating everything around the center is a continuous map without a fixed point). It is kind of amazing that the open problem Scott Garrabrant posted here requires (almost) no further tools to be formally stated than what is covered by the first two chapters of this book!

Perhaps the fundamental reason why my intuition is wrong is that we aren't trying to study nature and its messiness, but we are trying to figure out how to *design* systems, where we can achieve a very high degree of precision?

Chapter 3: Connectedness and Compactness

A space is **connected** if it can't be separated into two open sets. It is **compact** if every collection of open sets that covers it has a finite sub-collection that also covers it.

If (X, T) is a topological space and $A \subset X$, a **limit point** of A is a point such that for every $O \subset X$ with $x \in O$ intersects A (the point x is then 'arbitrarily close' to the rest of A; it might or might not lie in A). The set of A plus all of its limit points is denoted $\overline{A}$; it is the same as the intersection of all closed sets that contain A. If X is compact, then every infinite set in X has a limit point. In a metric space, the reverse is also true. Closed subsets of compact spaces are compact.

Connectedness and compactness are "topological properties", they are preserved under a homeomorphism. This fact can then be used to prove that $[0, 1]$ and $(0, 1)$ are not homeomorphic: the set $[0, 1]$ is compact but $(0, 1)$ isn't (the sequence $(\frac{1}{n})_{n \in N_+}$ has no limit point). Similarly, while they are both connected, you can take the point 0 or 1 out of $[0, 1]$ and it is still connected, but taking out any point of $(0, 1)$ leaves an unconnected space (and if there were a homeomorphism f between them, then $f : [0, 1] - \{p\} \to (0, 1) - \{f(p)\}$ would also be a homeomorphism). Connectedness can also be used to show that $[0, 1)$ and $(0, 1)$ aren't homeomorphic.

Theorem. Let (X, d) be a compact metric space. Let $f : X \to X$ be a map such that $d(f(x), f(y)) < d(x, y)$ for all $x \neq y$ in X. Then f has a unique fixed point.

Proof. This was difficult for me at the time. (Some adjustments made to readability, but not to the chain of arguments.) Is there a qualitatively shorter way? I don't know.

This is less powerful than the fixed point theorems for D^n because it demands that f has this property, but the upshot is that it works for every compact metric space.

Meta-insight for proving theorems: always have pen and paper, always make little drawings. It's low effort and almost always helps.

Chapter 4: Countability and Separation Axioms

Since the definition of a topological space is so general, there are a bunch of properties that feel useful but aren't always met. So mathematicians have defined them and given them names. Now, if one can prove that they are met, a number of useful results are immediate.

The **Hausdorff** property states that for any two different points x and y in a topological space (X, T), there exist open sets $O, U \in T$ such that $x \in O$ and $y \in U$ and $O \cap U = \emptyset$. **Regularity** demands the same (two disjoint open sets) for a point x and a closed set C; **normality** for two closed sets.

Theorem. Let (X, T) *be a topological space. If* X *is compact and Hausdorff, then* X *is normal.*

Proof. (Skit.) We first show that X is regular. Let $C \subset X$ be closed and let $x \in X - C$. For each $y \in C$, choose disjoint open sets U_y and O_y such that $x \in U_y$ and $y \in O_y$. The collection $\{O_y\}_{y \in C}$ covers C. Choose a finite subcollection $O_1, \ldots, O_n$ that also covers C (closed subsets of compact spaces are compact). Then $U_1 \cap \ldots \cap U_n$ is an open set around x and $O_1 \cup \ldots \cup O_n$ a disjoint open set around C. Thus, X is regular. To prove normality, given closed sets $C, D \subset X$, repeat the argument with open sets U_i around each $x \in C$ and disjoint open sets O_i around D (use regularity). //

Theorem. Let X *and* Y *be topological spaces, let* Y *be Hausdorff. Let* $f : X \to Y$ *be continuous. Then the graph of* f *defined by* $\Gamma_f = \{(x, f(x)) : x \in X\}$ *is closed in the product space* $(X \times Y)$*.*

Proof. (Skit.) I haven't defined the topology on a product space here, though.

I've done this proof twice, once when I worked through Munkres' book, and once as an exercise in the lecture on topology I've taken the past semester. My second proof (the one above) is much shorter and also simpler. Does that mean I improved?

There is another theorem which states that a topological space X is Hausdorff if and only if the **diagonal** $\Delta_X = \{(x, x) : x \in X)\}$ is closed in $X \times X$. In the lecture, this was given on the same sheet as the exercise above. And indeed, using the result above, one of the implications becomes a triviality: the identity map $\mathrm{id}_X : X \to X$ is continuous, so Γ_{id_X} is closed, and $\Gamma_{\mathrm{id}_X} = \Delta_X$. In the spirit of the lecture, it was considered stupid to prove this using primitive arguments. It's far simpler with the above theorem! But in

Munkres' book, it was an exercise in chapter 2, before functions were even introduced. And it was really difficult for me. Did the book waste my time?

I don't think so. The lecture tried to get away from primitive arguments as quickly as possible. Only use them if it is absolutely necessary, and optimize the structure of lecture and exercises for the ability to do everything as elegantly as possible. But why would that teach the right skillset? This has been on my mind a lot, and I think Munkres has the better idea. There is certainly a spectrum here, but optimizing for elegance only seems wrong.

Chapter 5: The Tychonoff Theorem

The Tychonoff theorem states that an arbitrary product of compact spaces is compact. The product topology is not the naively most obvious way to define a topology on a product space, however, and the result does not hold for product spaces in the box topology (although there are different & simpler reasons to prefer the product topology). The proof of this general result is far harder than the proof that finite products are compact.

The proof requires Zorn's Lemma**,** which is equivalent to the Axiom of Choice**,** which is the last axiom of ZFC (the "C" stands for "choice"). An alternative proof uses the Well-Ordering theorem, which is also equivalent to the Axiom of Choice.

Chapter 6: Metrization Theorems and Paracompactness

The metric topology is very well-behaved and understood. If one could prove about a topological space (X, T) that there is a metric d on X such that d induces T (in short, if (X, T) is **metrizable**), then one immediately gains a long list of useful properties that are met by X. This is why theorems that find conditions on a space which imply metrizability are of interest. The first such result proves that **regularity** (separation axiom) and having **a countable basis** (countability axiom) together imply metrizability. A stronger result weakens the requirement of having a countable basis and proves logical equivalence of metrizability and regularity & having a basis that is **countably locally finite**.

The concept of **local finiteness** sounds odd but turns out to be useful. A collection of subsets of X is locally finite if for every point $x \in X$ there is an open set around X which intersects only finitely many of them. The collection of all intervals $(n, n+1)$ is locally finite in R but is not finite. There are also local versions of the properties compactness, metrizability, connectedness, and **path-connectedness**. In the latter two cases, neither of the two versions (normal and local) imply the other.

Chapter 7: Complete Metric Spaces and Function Spaces

This was the most difficult chapter for me. I find it hard to deal with sets of functions – a function $f : R \to R$ can be thought of as a point in the infinite-dimensional space R^R, but how does one visualize an (open or otherwise) set of such points? It is made more complicated still by the fact that there are as many as four different topologies introduced on function spaces. The "normal" one, that is, the topology one gets from simply imposing the **product topology** on the space R^R corresponds to a sort of "point-wise" study of functions. In particular, a sequence $(f_n)_{n \in N}$ of functions converges to a function f in the product topology (convergence like continuity is a purely topological property) if and only if it converges point-wise (as defined in analysis). Similarly, it converges in the **uniform topology** if and only if it converges uniformly (as defined in analysis). Then there is the **topology of compact convergence** and the **compact-open topology.**

A metric space is called **complete** if every Cauchy sequence (= a sequence of points whose pairwise distances become and remain arbitrarily small) converges. The diameter of a set in a metric space is the supremum of pairwise distances in the set.

Theorem. A metric space (X, d) *is complete if and only if every sequence of closed nonempty sets* $A_0 \supset A_1 \supset \ldots$ *such that* $\mathrm{diam}(A_i) \to 0$ *has a nonempty intersection.*

Proof. *Skits.* This is one of those rudimentary proofs that I think are good practice. The drawings are both for the second direction of the proof.

Chapter 8: Baire Spaces and Dimension Theory

I've only started this and done a few exercises. The definition of a Baire space is very unnatural feeling and I don't yet have any intuition of why it is useful.

Chapter 9: The Fundamental Group

This probably takes the cake as the hardest chapter, but I had an easier time with it than with chapter 7, because I found it truly fascinating – unlike ch7, which felt like more of a grind.

A **path** on a topological space (X, T) is a continuous map $p : [0, 1] \to X$. The points $p(0)$ and $p(1)$ are called the *endpoints* of p, and p is said to go from $p(0)$ to $p(1)$. If $p(0) = a = p(1)$, then p is said to be a **loop** based at a. The space is path-connected if there exists a path from x to y for any $x, y \in X$.

There is an equivalence relation on the set of all loops based at a fixed point $b \in X$, where $p \sim q$ iff there is a **path homotopy** between them, which can be thought of as a continuous deforming of p into q such that the base point remains fixed. On a convex vector space, any two paths are path homotopic, since one can just connect them pointwise by a straight line. But on the circle S^1, the path $p : s \mapsto e^{2\pi is}$ that goes around the circle once is not path homotopic to the path $q(s) \equiv (1, 0)$ that just sits at a single point. The base point has to remain fixed, so there is simply no way to undo the one circulation. One can think about a rubber band wrapped once around a disc; at any point, the band can be pulled apart to make it longer, but without undoing the base point or leaving the circle, it can't be reduced to a path that does zero circulations (or more than one).

Two paths p and q where $p(1) = q(0)$ (such as two loops with the same base point) can be connected by simply going along p first and then q. The resulting path is denoted $p * q$. This operation can be proven to be well-defined on equivalence classes [p] of paths under homotopy equivalence. Furthermore, for any path p there exists a reverse path $\overline{p}$, and $[p * \overline{p}] = [c]$, where c is a constant path. And with that, the set $\pi_1(X, b)$ of all equivalence classes of loops based at b with operation $*$ forms a group! It is called -t-h-e- **a fundamental group** of the space X. Not 'the' because if X is not path connected, then fundamental groups at different base points may be different.

My favorite thing about this is that every continuous function $f : X \to Y$ with $f(b_0) = b_1$ defines a function $f_* : \pi_1(X, b_0) \to \pi_1(Y, b_1)$ via $f_*([p]) = [f \circ p]$, and the function f_* is a homomorphism between the two groups $\pi_1(X, b_0)$ and $\pi_1(Y, b_1)$. That means if f preserves the topology, then f_* preserves structure! And to make the analogy perfect: if f is a homeomorphism, then f_* is an isomorphism! And this is not only beautiful, but it also proves that the fundamental group is a topological invariant. Homeomorphic spaces have isomorphic fundamental groups, and the contrapositive statement is that if two spaces do not have isomorphic fundamental groups, then they are not homeomorphic. So the fundamental group is a way to prove that two spaces are 'topologically different'. It is more general than arguments based on the handful of topological properties studied previously, but not strictly more general; the spaces $[0, 1)$ and $[0, 1]$ both have the trivial fundamental group.

There's more. The fundamental group of the circle can be used to prove the fundamental theorem of algebra, which is pretty surprising and a strong knockdown to my concerns that theorems which care about single points can't be useful. It can also be used to prove that for any two bounded polygonal regions in R^2, there is a single cut that divides both exactly in half.

The fundamental group of S^1 is isomorphic to the infinite cyclic group $(Z, +)$. The homotopy classes are exactly determined by how often each path goes around the circle (and it can go around it in two ways, hence the negative numbers). The fundamental group of the sphere S^2 is not homeomorphic to $(Z^2, +)$, but to the trivial group $(\{1\}, *)$. There are also spaces that have non-trivial finite groups.

Chapter 10: Separation Theorems in the Plane

The last quarter of the book consists of much shorter chapters. I've only started this one, it (among other things) about how continuous maps $f : S^1 \to R^2$ always divide the plane into two regions, one bounded and the other unbounded.

Chapter 11: The Seifert-van Kampen Theorem

Chapter 12: Classification of Covering Spaces

Chapter 13: Classification of Surfaces

This was done extensively in the lecture, though in such a way that I felt like we didn't truly prove anything. This is a feeling I've never had reading this book!

Conclusion: This book is great. It's well structured, everything makes sense, everything is built neatly on top of each other, and the number of exercises leaves nothing to be desired. In general, I've had thoroughly positive experiences with Miri's guide; I've so far studied with four of the textbooks linked there, and all of them have been great (and the non-textbooks, too!). I don't do well learning out of source material that frustrate me (which happens a lot), so having a collection of high quality textbooks across a wide variety of topics has been extremely helpful. I'm planning to work through as much material as I can while I'm finishing my master's degree.

The worst thing I can say about this book is that it doesn't seem quite as impressive as *Linear Algebra Done Right* and *Computability and Logic.* In case of these two books (particularly the former), I've just been blown away by how much better and easier they are than my previous introductions to these topics. Nothing in this book gave me that impression, but as I said, it is still extremely solid. And it should be said that it covers a much larger and more difficult subject.

To perform best at work, look at Time & Energy account balance

Several weeks ago, I got a chance to join a talk hosting one of the very few female regional head at Google.

Despite not having any business background, she climbed the rank from entry level employee to become a regional head, surpassing everyone else from prestigious business degrees and rich experiences.

One success driver she mentioned got my attention. Despite lagging very much behind at the beginning, the core to her success is that **she always aims for 120% result of any task in front of her.**

The reason why this interests me is not because of my fresh ears.

In fact, this is not the first time I heard of this concept. Not the first time I get inspired of giving it all to whatever is in front. Not the first time I try…and not the first time I fail.

Did I not put in enough effort?

No…in fact, I put in so much effort to make this concept come to live, not realising that while effort is highly important, it’s critically inadequate.

As I listened to this amazing regional head talking about different aspects of her life, I came to realisation on what I have always been missing so far.

To make each task yield 120%, apart from effort, we should also look at our **time and energy balance.**

Contributing the best on a task means to give the amount of time and energy in the level required to make the result best.

We cannot contribute what we don’t have.

No matter how much effort we try to give adequate time required for the best, we only have 24 hours a day.

No matter how much energy we try to put into each task, we only have a limited stream in each day.

Therefore, giving our best does not start from the moment we begin working…but from the moment we plan our schedule and project pipelines.

When having "Enough Time" is Not Enough When my boss asked if I have enough time to take on one additional project, I would look at how much time is required to finish all the tasks on my desk and then, most of the time, said "Yes" thinking I have enough time to finish it all.

However, there is a difference between *having enough time to finish it all* and *having time to make it best.*

Coming back to evaluate all the projects in my pipeline, I realize that the time I have is only enough to finish all up, but not to go above and beyond.

I have two choices:

- Finishing a lot of tasks with average results OR
- Complete major task with the best impact that goes beyond expectations

There is no right answer here, but for my situation, the second works better.

Even having Time is Sometimes Not Enough Having time is good. But having time without full energy...hmm...unlikely to be productive.

Another good lesson I learned from this talk is that *ample time to do it best* should always come with *ample energy*.

It's just normal to plan business projects with the right balance between high-low energy requirement. However...our energy pool is not limited only in working hours, but also in personal life.

One thing I learned is that when looking at high-low energy requirement in my activities list, I should include all activities both in office and at home.

Despite saying "I only have one major project going on during working hours", if this lady has to practice running a marathon at night with high intensity, how would she have enough energy to do both best, despite marathon not being related to works.

To summarize, with one key success driver in career being to do our best in the tasks at hand (eg. the concept to deliver 120%), many ambitious people try to put in so much effort to ensure the best results. However, the best results actually begin even before we start doing each task...but begins during project planning, in which *time and energy balance* would determine how our project results would turn out to be.

Active Curiosity vs Open Curiosity

I think the word ‘curiosity’ is used to describe two distinct things that I will now differentiate as active curiosity and open curiosity.

Active curiosity is driven & purposeful. Like thirst, it seeks to be quenched.

When you see a blurry object among distant waves and it looks like it might be a humpback whale, and you *want to know*.

When you are asked a trivia question like, “How many people have seen the broadway show *Hamilton* more than once?” or “What’s the life expectancy of people my age in the US, in 2019?” And you find yourself wanting to go to Google.

When you watch a YouTube video of someone doing something crazy, and you’re like, *How did they DO that?*

When you hear someone mention your name from across the room, and you become anxious to know what they’re saying about you.

Active curiosity activates the part of your brain that anticipates a reward, and it can enhance learning, making it easier to remember surprising results. [1, 2]

//

There’s another kind of curiosity that is often referred to by therapy books and practitioners. It is phenomenologically different, and it seems good to be able to distinguish the two types.

This type of curiosity, which I’ll refer to as **open curiosity**, is best achieved when you feel safe, relaxed, and peaceful. In my experience, it basically requires parasympathetic nervous system activation.

I’m aware of at least one person who can’t recall experiencing this type of curiosity. So I don’t expect this to be a common or universal experience, but I think it’s achievable by all human minds in theory.

This type of curiosity isn’t very driven. It doesn’t need satisfaction or answers. It is open to any possibility and can look without judgment, evaluation, worry, or anxiety.

It is evoked by the Litany of Gendlin and the Litany of Tarski. It is related to original seeing / boggling / seeing with fresh eyes.

When I have open curiosity, I do have things I’m curious *about*! So it isn’t a totally passive experience. I often use open curiosity to get curious about myself or another person. It’s a very useful state for doing therapy-related work, as all emotions and thoughts feel acceptable and manageable, rather than overwhelming or undesirable.

Perhaps strangely, this type of curiosity is open to knowing, in addition to not knowing. It is open to understanding, in addition to not understanding. It doesn’t *need* to know or understand things, and as such, you can sit with confusing, upsetting, or vague things. And you can just ask questions about them, with an open mind, ready for whatever response or reaction comes. If no answer comes, it doesn't feel like a problem. You can just ask another question.

I don't recommend using open curiosity to study for your exams or read *Superintelligence* or learn how to make things. It's not good for downloading lots of new information or developing a skill. Active curiosity is what you want for that.

I do recommend it for the following:

- Introspection
- Holding space for a friend who's upset / has a lot of feelings
- Trying to resolve a heated conflict that you're involved in or mediating
- Understanding how you relate to things like death, insanity, suffering
- Creating an intimate moment with someone
- Watching weird, surreal, artsy movies
- Being in nature or somewhere very unfamiliar
- Circling, meditating, therapy, IDC, etc.
- Gaining insight into the universe, yourself, etc.

When I try to use *active curiosity* to understand how a person's mind works, they often feel examined under a microscope, like they're an experiment on my surgical table. When I try to use active curiosity to watch an artsy movie, I feel frustrated that it doesn't make any sense. When I try to use active curiosity when my friend is upset about something, they feel unheard and like I'm just trying to fix their problem to make it go away; I also tend to ask unhelpful questions (more selfish interest in understanding the situation / update my opinions than trying to help them).

//

Now that I've described these two types: Do they resonate with you at all? Do you basically know what I'm talking about, and it's crystal clear? Or does this seem confusing and alien? I find it quite easy to distinguish the two in myself, and I wonder if others feel the same.

(It also seems very plausible this distinction is already covered in research literature or even on LessWrong, and I just didn't look very hard! References welcome.)

I would like to start using these terms to be less vague when I talk about "curiosity."

I notice I try to talk to certain people based on which type of curiosity I expect from them. Sometimes, I want active curiosity, like when I'm trying to think through a concrete problem or I want their opinion or advice. Other times, I want open curiosity, like when I'm having emotions, going through a confusing situation, or want to feel heard or accepted.

I have a list of people I can rely on for active curiosity; and a separate list of people I can rely on for open curiosity. (These lists don't really overlap?)

But I haven't really tried to just ASK for one type or another from someone.

Now that I've named the types, maybe it will be easier to refer to which one I'm wanting, and people can help by saying which one they can potentially offer.

(For the record, if you want open curiosity from me, this is something I can usually switch on, especially on a good day. If you want active curiosity, it depends more on the topic of the conversation and on the object-level details, so you may want to tell me what the subject matter is first.)

The tech left behind

Hello, I am asking for some insights for a research I am doing. Can you cite examples of technologies that have been forgotten? What I mean by "forgotten" is not things we don't know how to do but used to (I suspect there aren't that many), nor things that are no longer in use but used to (mechanical television), but things that were decently developed (either in theory or in practice) but never "saw the light of day" anyway.

It's my first time posting, so I won't do much policing on the answers, thanks in advance.

February gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2019/02

Plans are Recursive & Why This is Important

Epistemic status: Reference. Highly confident. I rely on what is presented here extensively in my own thinking.

Plans are recursive. Any plan can be decomposed into parts and those parts can be in turn be decomposed into further parts and so on until it is senseless to decompose any further. This is not a profound point, it is something we all appreciate intuitively and explicitly to at least some degree or another.

Nonetheless, the recursive nature is of such fundamental importance to the entire practice of planning that it warrants an explicit treatment.

- Crucial principles of good planning are readily derived from a recursive model of planning. Having a solid grasp of the recursive models makes it harder to forget these key principles.
- An explicit treatment of a topic can help it sink in deeper to intuition even if one already has some sense of it.
- Even if everyone has an intuitive sense of something, it can be hard to talk about the thing if there isn't a common explicit handle.

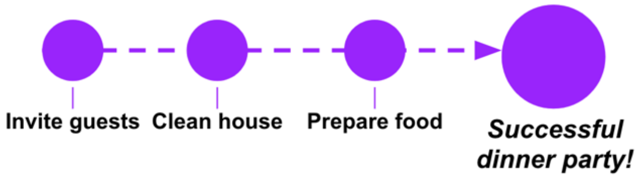
A quick definition of *planning* as used in this post: *Planning is the selection of actions to be deliberately executed in order to achieve a desired outcome/goal.* [1]

Plans are recursive

Usually, we treat plans in a very linear fashion. Step 1, Step 2, Step 3. To-Do lists capture this, and we can also draw an accompany thing diagram.

Example Linear Plan: Hosting a Dinner Party

1. Invite guests
2. Clean house
3. Prepare food



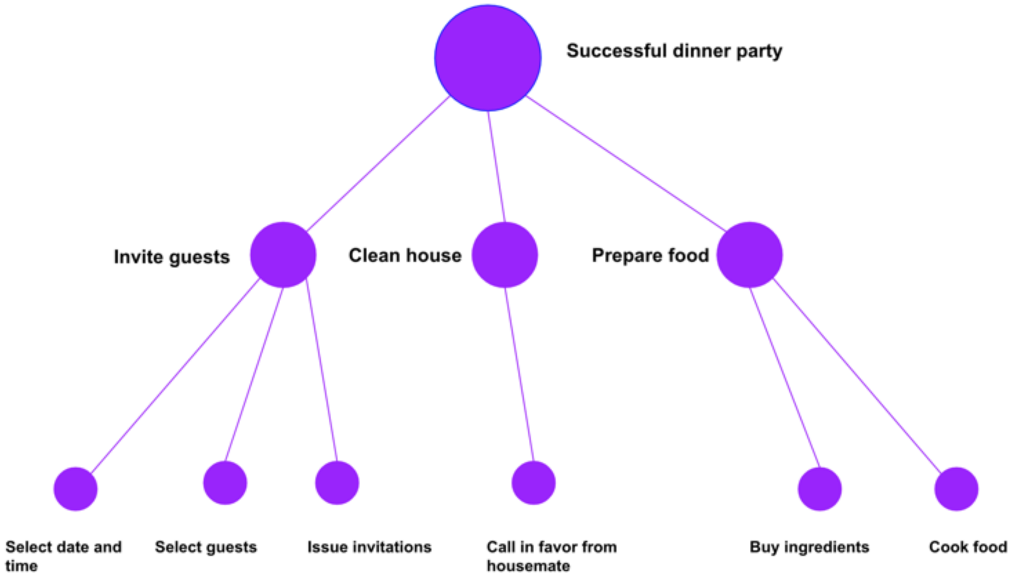
Linear diagram of a plan to host a dinner party

Of course, we all recognize that the three steps of the plan above are composed of further sub-steps. Nested to-do lists capture this well.

Example Recursive Plan: Hosting a Dinner Plan Expanded

1. Invite guests
   1. Select date and time
   2. Select guests
   3. Issue invitations
2. Clean house
   1. Call in favor from housemate
3. Prepare food
   1. Buy ingredients
   2. Cook food

These steps could be further decomposed, but this suffices to demonstrate the idea. We can draw another diagram.



Recursive tree diagram of a plan to host a dinner party

Why call it recursion?

Sometimes people use fancy technical terms in order to sound smart even though a more commonplace word would suffice. Arguably, I could have said that plans are hierarchical or nested or something. Why use the mathematics and computer science term, *recursion*?

I defend this choice with reference to the defining features of recursion:

1. You solve a problem by breaking it into smaller instances of the same type.
2. Each time you decompose the problem into a smaller problem, you apply the same function/procedure to the smaller piece as you did the larger one.
3. You continue decomposing until you reach a base case which requires no further decomposition.

4. The overall problem is solved by rolling up all the solutions to lower-level decomposed sub-problems.

[The calculation of factorials is the canonical introductory example for recursion and neatly illustrates all of these features.]

Planning has all of these features and is therefore rightly described as being recursive.

We see that plans can be decomposed into sub-plans which might then be decomposed into further sub-plans. Eventually you stop, i.e. you reach a base case. When is this? It depends on the context, but one heuristic is to stop when the level is so trivial that it doesn't require further decomposition. In the above example of hosting a dinner party, I might not need to recurse to the level of planning out the steps of sending an emails. I can take for granted that I know how to carry that out.

The Core Planning Process

What is the universal planning function/procedure in the context of planning? A decision theorist or AI expert might be able to offer a little more rigor here, but the universal core planning process we repeatedly apply whenever we are planning is going to look roughly like this:

- Enumerate the set of possible actions to take.
- Predict which actions will result in which outcomes (and with which likelihood).
- Assign relative preferences to each of the potential outcomes.
- Use the above points to prioritize the actions you will take based on the expected costs, benefits, and risks associated

How the human brain executes the above steps is complicated. And what is required to execute those steps will also differ dramatically between plans and at different levels within a plan. In particular, sub-plans can require very different models to map actions to outcomes than the models used at a higher level. The models used for business strategy are different from those used to file taxes, though both are part of the broader "business success" goal. Still, we shouldn't let the variation between plans mask that the core steps are always the same.

Plans are Two-Fold Recursive

We can actually say that planning is doubly recursive. Planning is both recursive in planning, i.e. when you're choosing what to do [2], and also in *execution*. To execute a plan, you must execute each of each the sub-plans, and each of the sub-plans' sub-plans and so on. When we're talking about execution, the recursive function to be applied at each level is simply "Do X".

Implications of the Recursive Nature

The following section details some key principles of good planning that can be directly derived from the recursive nature of plans. Even though both the recursive nature and the principles presented here seem sensible, if not obvious, on their own, linking the two together makes it clear why these planning principles are necessarily so.

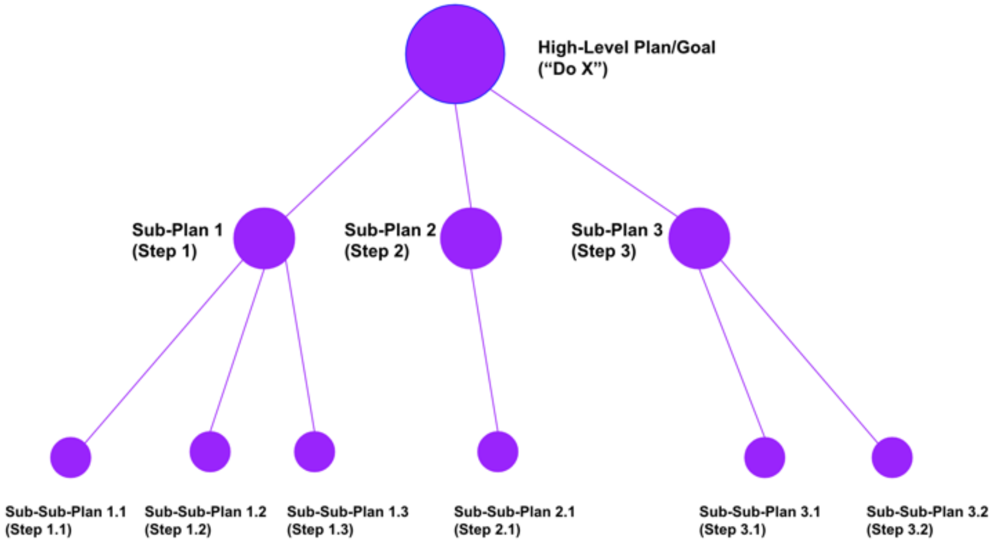
Goals/Steps/Actions/Plans are Isomorphic

When we plan, we usually start with a goal. Then we figure out actions and sub-plans we will take towards the goal. This picture is somewhat misleading. It makes it sound like the goal and steps we take towards it are of a different kind, and so encourages us to treat them differently. Usually, that means treating the goal as more fixed than it ought to be.

Someone might decide that they have the goal of becoming a lawyer. They start of with this goal and start evaluating steps they might take towards achieving it. Maybe they start thinking about which law school they want to go, where they will get a student loan, how they're going to improve their study habits. They start considering the obstacles and how to overcome them.

For each possible step towards the goal, they enumerate and evaluate the options for their benefits, costs, and risks. What is so easily missed is whether or not the "goal" was the right step to begin with. One must remember that the goal itself is likely only a step within some yet larger step. **The goal is a step too because, from the perspective of the recursive tree, all the nodes are of the same type.** [Perhaps we wish to treat nodes at the very top or very bottom differently, but otherwise, there's no difference.]

It is only in our minds that we distinguish between goal, step, action, and plan. Structurally they are the same thing.



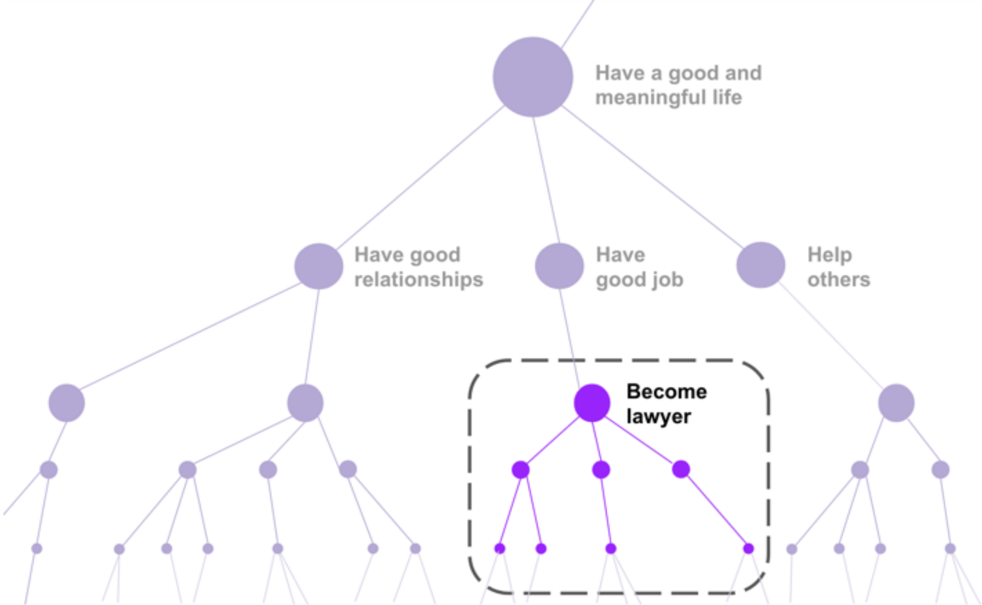
See? They're all just plans.

The exception to the isomorphism are *terminal goals* which we wish to accomplish purely for their own sake rather than *instrumental goals* which we seek to accomplish to attain something else. Terminal goals are not steps within another plan and so the isomorphism does not hold with them. Still, I claim that since the overwhelming majority of our goals and plans are only instrumental, it is reasonable to assert that goals (generically) are isomorphic to plans. Most nodes in the recursive tree are in the middle. It's true enough that it's helpful to remember it as such.

Almost All Plans are Sub-Plans; Almost All Goals are Sub-Goals

The corollary to the above point that goals/steps/actions/plans are all the same is that [almost] all plans are sub-plans and [almost] all goals are sub-goals. If you have a goal, it’s probably because there’s something you hope that goal will accomplish.

We can reuse the lawyer example from above.



What can be lost when you start with your goal and don’t look any higher is all the reasons you selected your goal in the first place. It is worth remembering and examining those reasons. Possibly you selected those reasons long ago and they no longer apply. Possibly it was an intuitive or emotional decision but not an especially good one. If you don’t think about the reasons behind your goals, i.e., your higher-level goals, then you very likely will either select the wrong goals or select the right approximate goal which you then execute in such a way as to not accomplish your higher-level goals.

Peter Thiel, famous billionaire investor, went to Stanford law school and practiced in a prestigious New York law firm for seven months and three days. He is oft quoted as describing that this law firm was a place where *“from the outside everyone wanted to get in and on the inside everybody wanted to get outside.”* When he left, a colleague told Thiel that he didn’t know it was "possible to leave Alcatraz." Clearly not somewhere he actually wanted to be. Peter Thiel cites his reason for going to law school in the first place: "*So I think I got a JD because of the abstract prestige, because I didn't know what else to do, and I didn't concretely think through what it would be like to be a lawyer.*"

Peter Thiel’s example highlights an important point. If you know why you have a particular goal, you can list out the assumptions and beliefs that cause you to have it. Once listed, the assumptions and beliefs can be examined, further evidence can be collected, and you can avoid making plans based on shoddy beliefs and limited information. You can ensure that you think through concretely what it would be like to be a lawyer, or as Thiel advises others, go identify someone with a JD who's a good role model.

The elucidation and examination of assumptions should be part of your "core planning process" which should carried out appropriately at each level and node of your plan.

In a sense, you need to dissolve the distinction between plan/goal/action/step and see the recursive tree of your plan for what it is. Then you apply your planning process to the nodes indiscriminately. When starting with a goal, you ought to "recurse upwards" until you identify higher-level goals that you are sufficiently confident in [3]. Even if you are completely certain about your higher-level goals, you still want to be thinking about them because their details dictate what the details of your lower-level goals need to be.

When our lawyer-wannabe sits down to plan how they're going to become a lawyer, they will make worse choices if they forget what the point of becoming a lawyer was. The point of a lawyer is not to become a lawyer inherently, more likely it's to have a good job with a high salary, prestige, and satisfying work as part of overall having a good life. You can avoid repeating Peter Thiel's mistake.

1. If you remember why you care about a goal, you can examine whether this goal will actually achieve your higher level goals. The lawyer-wannabe can look up lawyer job opening and salaries, find lawyers she can question about their lifestyle, poll some friends to see if they actually respect lawyers, and so on. She might find out being a lawyer doesn't serve her higher-level goals and should maybe pursue something else. It'd save her a lot of time to find that out sooner rather than later.
2. It's possible that there are versions of becoming a lawyer which meet the higher level goals and versions which don't. Becoming a barrister might have a lot of prestige and high salary yet be stressful and have extremely long hours. Conversely, being an environmental lawyer might pay less but an overall better lifestyle. Keeping the true goal in mind helps determine the details you want to target.

I fear that the very particular obvious career example I have presented won't translate readily in the minds of readers to all the other cases where this applies. It applies to projects within startups and business, it applies to the books you read, it applies to the people you date and how you find them, it applies to your hobbies and your vacation. It applies to any planning of non-trivial complexity. I encourage the reader to search for a few examples from their own life where they might benefit from applying this line of thinking.

A goal is only right if it actually serves your higher-level goals

If you didn't pick the right goal to begin with, then no matter how hard you hit that goal, all is for naught. In terms of a recursive trees, for a goal to matter, there needs to be an "unbroken path" to something you terminally value.

If the lawyer-wannabe succeeds super hard at becoming a lawyer: graduates Summa Cum Laude from the best school, gets a role in the most prestigious firm, etc., it doesn't matter at all if it turns out being a lawyer was the wrong choice.

To reword points from above, goals can be wrong in two ways:

1. You picked the wrong goal entirely. Another option was entirely better.
2. You picked roughly the right goal, but you implemented it such that it had no value.

An aside about long-term planning and feedback

High-level goal selection is often fraught because there is slow feedback. A young person might not discover for years that they didn't actually want to be a lawyer, so it's easy to pick the wrong high-level goal and grind towards for years, only getting feedback that it was the wrong choice after they graduate and start working. Peter Thiel is an example. Another class of examples is startups and business deciding which products to build. Done wrong, a startup can pick a long-term goal and work for months or years before finding out it was the wrong goal.

Worse than the pure absence of quick-feedback, long-term planning is often influenced by short-term factors such as emotions. "Gah, planning is stressful! I'm just going to pick something and stick with it." Rather than focus on the long-term goal, how the person feels in the moment is prioritized. Relatedly, as with Peter Thiel, instead of trying to pick the best long-term plan, someone might pick the plan which gets the greatest social reward (since there are indeed rewards for just having cool plans). There can faster feedback on whether your plan is cool then whether it will succeed. You'll know as soon as you boast to your friends or describe your product plans to your startup's investors. The quick feedback on emotions and social rewards often outweighs the value of a long-term goal one is supposedly planning towards. Beware.

The precise success criteria of a lower-level goal are determined by the higher-level goals

The importance of being clear on why a goal has been selected is never clearer than in the case of delegation. A manager asks a research engineer to write a report on their recent research. The engineer does so, but the manager is not pleased! The report is a technical, suitable only for other engineers, yet this is not what the manager wanted - the manager wanted a report to give to the business executives.

The research engineer delivered what was requested "a report on their recent research", but since this didn't connect with the higher level goal of having a way to update the executives, it was worthless.

The danger is that we humans often encapsulate our goals and plans very crudely and with very little detail. A few words like "become a lawyer." We rely on ourselves and others implicitly understanding the specific versions of those goals which would count as success - yet so often we don't.

Even when the goal is certain, you need to understand the bigger picture in order to be able to make lower-level implementation choices. Anyone delegating a task is advised to instruct the delegatee in why the task is being requested. The more detail the better.

You can't just ask someone to book you a venue for an event. You either need to tell the exact kind of venue you want or tell them what kind of event it is so they know what kind is appropriate. And even if you say *“fits 50 people, within X budget, within Y miles of location A”* there could still be further decisions whose ideal selection is made only if you know *why* a venue is needed: what kind of event, which guests and therefore which ambiance and so on would best.

Delegation is hard because it means splitting the recursive tree among multiple people. In the case of principal-agent problems, you’re transplanting a goal from one person’s tree to another. You can see why that’d run into trouble.

Final example: as a Data Scientist, I was often asked to build things with little specification and no explanation of why. A technical example here is that if you don’t understand how a classifier will be used, then even if you can build one, you don’t know how to make a dozen relevant design choices, e.g. setting thresholds which determine false positives and false negatives. Building something without an idea of why is a great recipe for building a version of it which doesn't meet the true user requirements. [4]

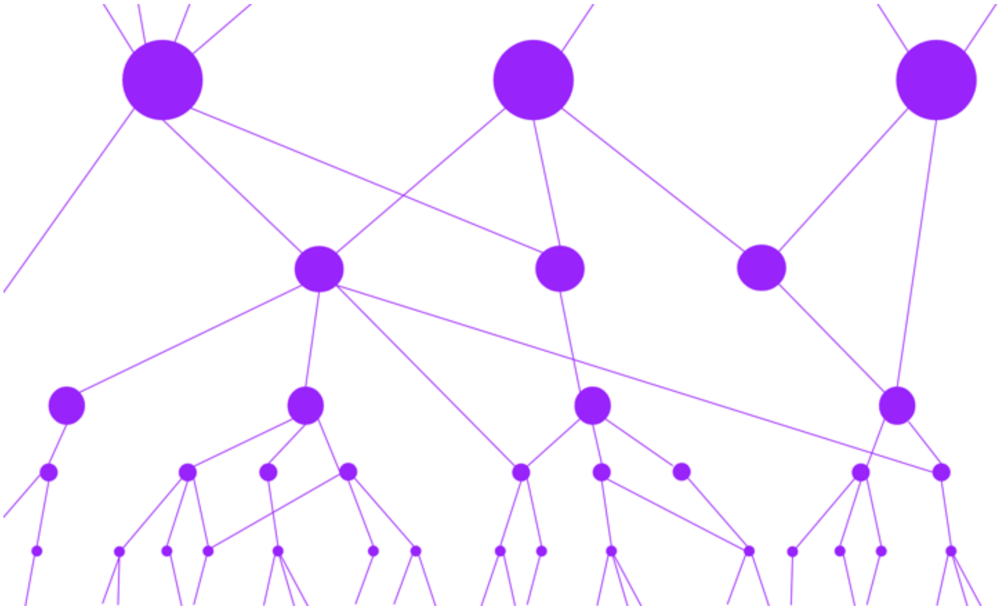
Wrapping up

Though the ideas presented here are simple, I hope that if you have read this far, this post will have deepened your appreciation of this fundamental property of plans and will result in you making better plans towards good things.

Appendix: Multiple Goals

Though it was skipped for the sake of simplicity in the main text, realistically, many our plans are executed with more than one goal in mind. This is entirely compatible with a recursive model of plans, but it makes the diagrams messier. In order to not distract from the main points, examples in the text were chosen to roll-up neatly into a single goal.

Below is a realistic recursive plan diagram illustrating multiple goals. Makes you appreciate what your brain has to deal with.



Modern art: *a recursive tree diagram of a plan where sub-plans feed into multiple higher-level plans, e.g. actions are taken towards multiple goals.*

Endnotes

[1] This broad definition stands in opposition to a common definition which uses planning primarily in contexts of scheduling, e.g. plan your day, plan your week. The broader definition here is essentially synonymous with decision-making, perhaps differing only in connotation. Decisions somewhat imply a one-off choice between options whereas planning implies selection of multiple actions to be taken over time. The term planning also somewhat more than decision-making highlights that there is a goal one wishes to achieve.

[2] With the broad definition of planning used here, it’s important to note that planning needn’t happen all at once. More commonly, different components of a plan (selection of action) happen at different times. Very commonly, higher-levels of the plans are selected up front and lower-level actions are selected near or at the time of execution.

[3] This is often really, very hard and time-consuming. If you’re haven’t recursed explicitly upwards to determine what you want and what your true high-level goals are, this might take a while including both the time to do it and possibly the time spent gaining the skills do it. The good news is that after initial investment you get to reuse your knowledge of your high-level goals repeatedly when making lower-level plans such that you make much better low-level plans.

[4] My boast: determined to not build the wrong thing, I spent enough time figuring out what the why reasons for our engineering work should be that my company eventually asked me to take that over as my explicit role.

Preferences in subpieces of hierarchical systems

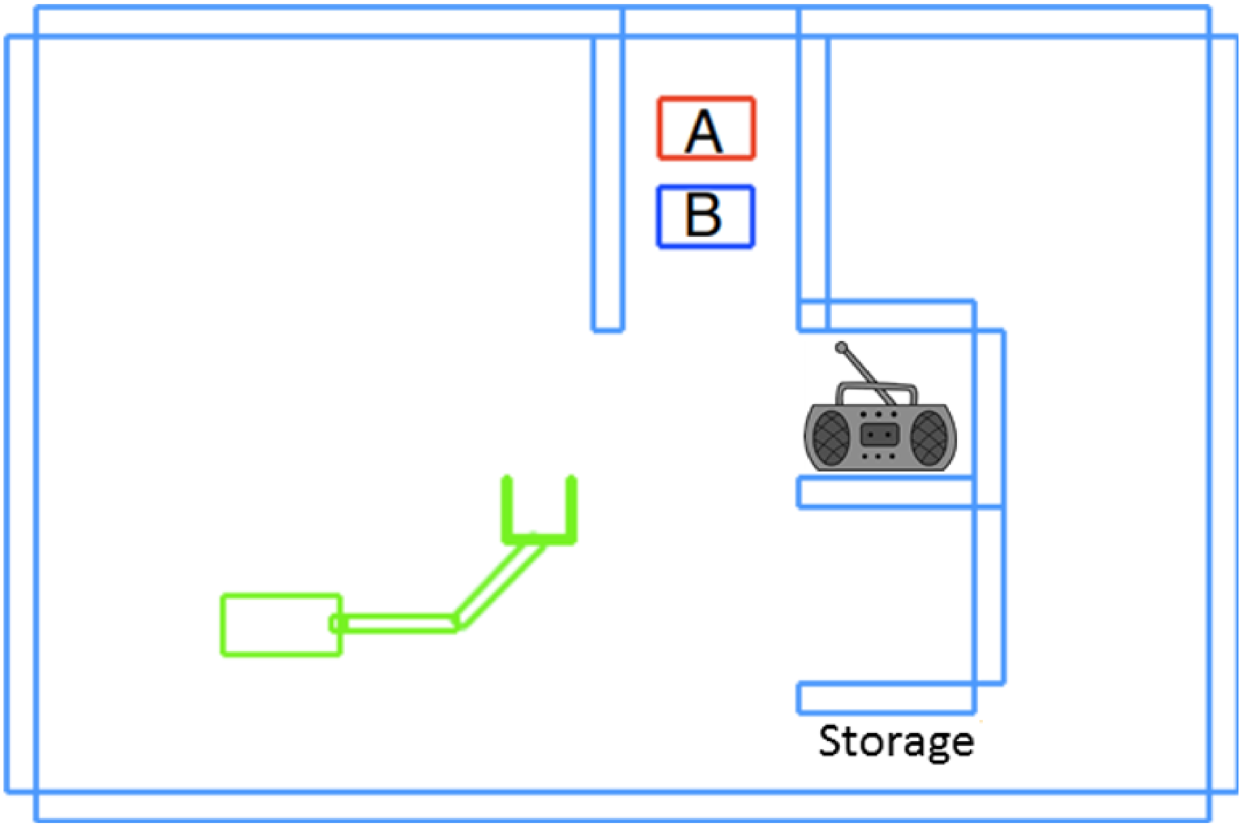
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

In a previous post, I looked at hierarchical systems, and at subagents within them that could develop their own preferences. This was due to a conversation with Dylan Hadfield-Menell.

This post will ignore subagents, and answer Dylan's original question: can we deduce the goals of a hierarchical system? So, we assume the subpieces of the system are not subagents worthy of ethical consideration. Now, given that the subpieces could be inefficient at "what they're supposed to be doing", how can we assume the system itself has a goal?

Algorithms and symbols, again

The situation is as it was in the initial post, a simplified version of the task in this paper. Here an agent is moving object A to storage:



Within the hierarchy of operations needed to move object A, there is a subalgorithm tasked with gripping object B; the code for that is as follows (this is a slight variant of

Algorithm 1 in the earlier post):

```
Algorithm 3:
State: S={Gripped(B)=False}
Goal: G={Gripped(B)=True}
Suggested plans, generated by SPG(S, G):
    (Radio_On AND Grip(B))=> Gripped(B)=True
While G=False:
    Run Suggested plans according to criteria C(G)
```

Here the criteria C(G) just assess whether the plan will move the situation closer to the goal G.

Let's look at the subalgorithm that is generating the suggested plans, which I labelled SPG (suggested plan generator). It takes the State S and Goal G as inputs.

It is obviously bad at its job: turning on the radio is not necessary or useful for gripping B. It's not utterly incompetent at its job, though: turning on the radio and then gripping B, will at least result in B getting gripped.

The job of an algorithm is to do its job

I've been talking about SPG's 'job'. But how do I know what this is (especially given that the term is probably undefinable in theory)?

Well, there are a number of indications. First of all, akin to "figuring out what Alice wants", we have the evocative names of the various variables: State, Goal, Suggested plans. If these labels are grounded, then they clearly indicate what the intended task of SPG is.

Note also that there is a loop where the plans of SPG are run until G is true. Even if the symbols are ungrounded, and even if SPG were a black-box, this indicates that fulfilling the the goal G is the purpose of Algorithm 3, and that SPG contributes to this. In this loop, we also have the assessment criteria C(G), checking whether the plan will move closer to G or not; given some weak assumptions, this is also an indication that G is a goal of Algorithm 3, and that coming up with plans that reach G is the job of SPG.

Bad job, unoptimiser

The previous criteria can establish what the job of SPG is, but doesn't allow us to say that it's bad at its job. It seems to be bad because turning on the music is an unnecessary step.

But imagine (situation 1) we're now looking higher in the algorithm hierarchy, and there is a cost function that counts how many steps are used to achieve the goals. Then we can say that SPG is doing a bad job; but the full criteria of that the system wants have not been passed down to the Algorithm 3 and SPG level.

Conversely, imagine (situation 2) we're looking higher in the algorithm hierarchy, and there is not cost function, but there is a desire for the radio to be on. Then SPG is doing a good job, even though the full criteria have not been passed down.

Especially if the system is capable of self-modification, we shouldn't expect all the job criteria to be located close to the subsystem itself. It's possible that a cost-assessor (in situation 1) has tried to analyse the whole system, and deemed SPG's inefficiency to be minor. Or, conversely, that a radio-turn-on assessor (in situation 2) has analysed the whole system, noticed SPG's behaviour, and let it be (or even added it in), because this helps achieve the systems's overall goal.

The general case

So in general, the role of a subroutine in a hierarchical system is to achieve the task that whatever called that subroutine wants it to achieve. The nature of this task can be be inferred by looking at grounded symbols, and/or at the structure of the algorithm that calls the subroutine, including what it does with its output. Some goals may be implicit, and handed down from higher in the algorithm's hierarchy.

Note that if the subroutine takes actions - either in the real world or by modifying global variables within the algorithm - these can also be used to define its task, especially if the global variables are grounded or the meaning of the actions are clear.

Better implementation

Once the role of all subroutines is established, the goal of the whole system can be estimated. Again, grounded variables are useful, as are the purposes that can be inferred by the fact that the system calls a certain subroutine (with a certain role) at a certain point.

Then once this goal is established, we can talk about how the system might improve itself, by making the outcome more in-line with the goal. But we can't talk about improvements in an abstract sense, without establishing this goal first. Even seemingly useless parts of the system, may be there deliberately.

More structure and information

The more structure a system has, the easier, in general, it is to assess its goal. If there are top-level subroutines that go around assessing the lower levels, then the assessment criteria are major components of the system's goal.

However, it might be that the system doesn't give us enough information to figure out its goal, or that multiple goals are compatible with it (this is most strongly the case if the systems variables are poorly grounded). This is to be expected; we can't infer the goals of a general agent. In this case, we are allowing some assumptions about grounded symbols, internal models, and hierarchical structure, to cut down on the space of possible goals. This might not always be enough.