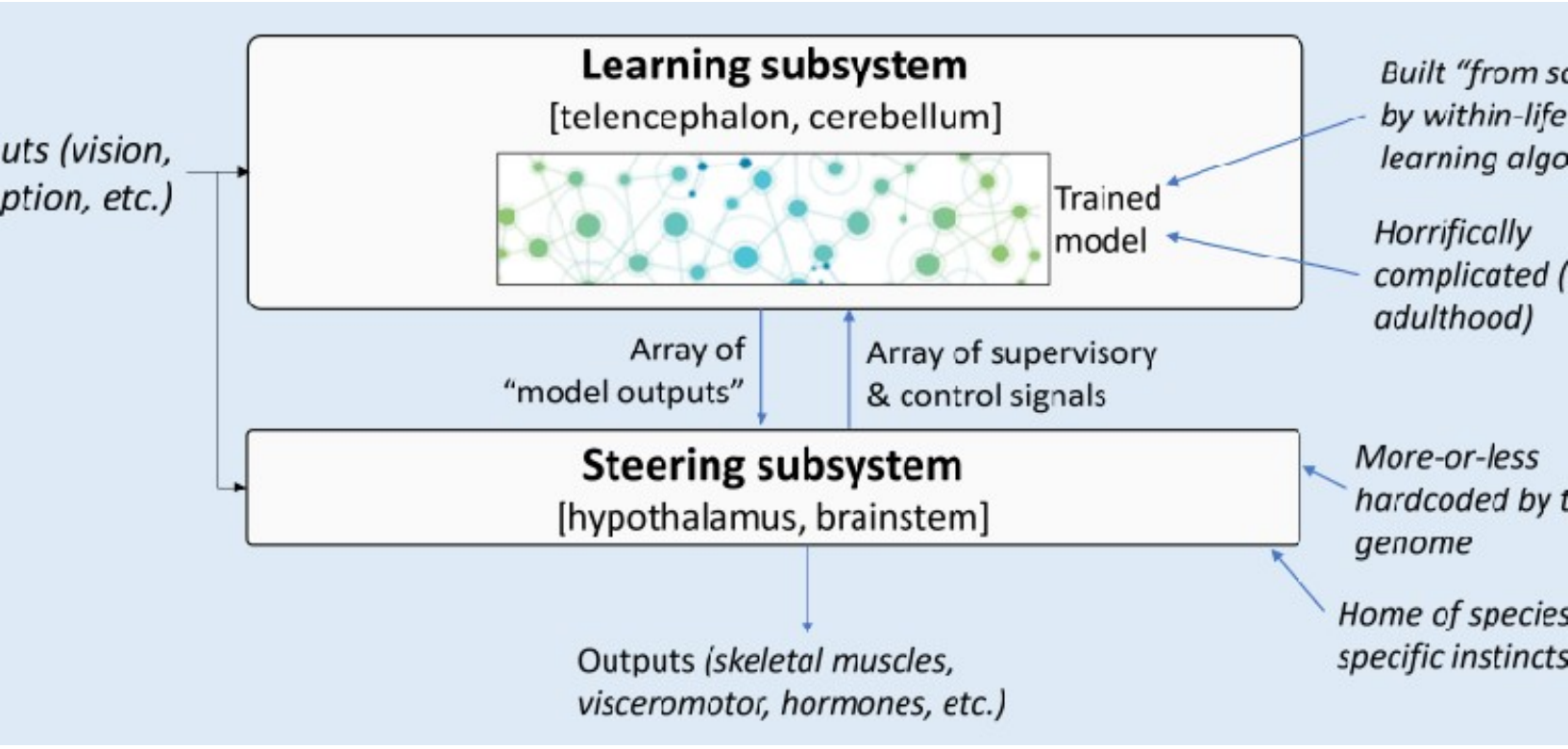

Learning subsystem
[telencephalon, cerebellum]
Trained model
Built "from sc
by within-life
learning algo
Horrifically
complicated (
adulthood)
uts (vision,
ption, etc.)
Array of
"model outputs"
Array of supervisory
& control signals
Steering subsystem
[hypothalamus, brainstem]
More-or-less
hardcoded by t
genome
Home of species
specific instincts
Outputs (skeletal muscles,
visceromotor, hormones, etc.)

Intro to Brain-Like-AGI Safety

1. [Intro to brain-like-AGI safety] 1. What's the problem & Why work on it now?
2. [Intro to brain-like-AGI safety] 2. “Learning from scratch” in the brain
3. [Intro to brain-like-AGI safety] 3. Two subsystems: Learning & Steering
4. [Intro to brain-like-AGI safety] 4. The “short-term predictor”
5. [Intro to brain-like-AGI safety] 5. The “long-term predictor”, and TD learning
6. [Intro to brain-like-AGI safety] 6. Big picture of motivation, decision-making, and RL
7. [Intro to brain-like-AGI safety] 7. From hardcoded drives to foresighted plans: A worked example
8. [Intro to brain-like-AGI safety] 8. Takeaways from neuro 1/2: On AGI development
9. [Intro to brain-like-AGI safety] 9. Takeaways from neuro 2/2: On AGI motivation
10. [Intro to brain-like-AGI safety] 10. The alignment problem
11. [Intro to brain-like-AGI safety] 11. Safety ≠ alignment (but they’re close!)
12. [Intro to brain-like-AGI safety] 12. Two paths forward: “Controlled AGI” and “Social-instinct AGI”
13. [Intro to brain-like-AGI safety] 13. Symbol grounding & human social instincts
14. [Intro to brain-like-AGI safety] 14. Controlled AGI
15. [Intro to brain-like-AGI safety] 15. Conclusion: Open problems, how to help, AMA

[Intro to brain-like-AGI safety] 1. What's the problem & Why work on it now?

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

1.1 Post summary / Table of contents

This is the first of a series of blog posts on the technical safety problem for hypothetical future brain-like Artificial General Intelligence (AGI) systems. So my immediate priority here is saying *what the heck is "the technical safety problem for brain-like AGI" and what do those words even mean and why on earth should I care*.

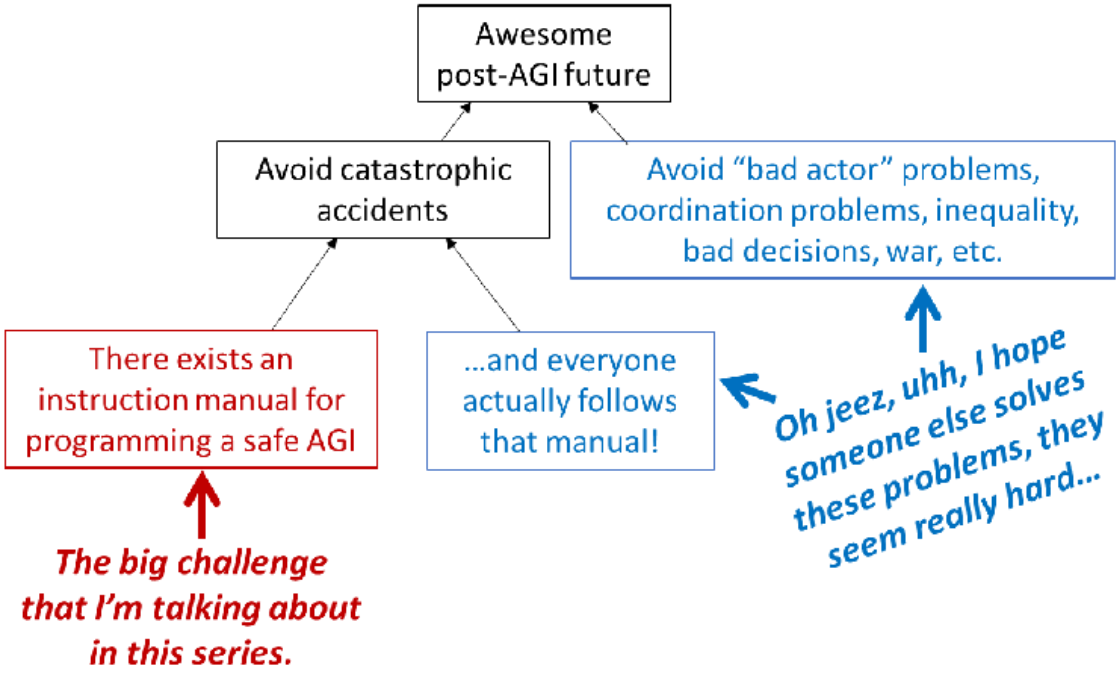
Summary of this first post:

- In Section 1.2, I define the "AGI technical safety problem", put it in the context of other types of safety research (e.g. inventing passively-safe nuclear power plant designs), and relate it to the bigger picture of what it will take for AGI to realize its potential benefits to humanity.
- In Section 1.3, I define "brain-like AGI" as algorithms with big-picture similarity to key ingredients of human intelligence, presumably (though not necessarily) as a result of future people reverse-engineering those aspects of the human brain. What *exactly* that means will be clearer in future posts. I will also bring up the counterintuitive idea that "brain-like AGI" can (and probably will) have *radically nonhuman motivations*. I won't explain that here, but I'll finish that story by the end of Post #3.
- In Section 1.4, I define the term "AGI", as I'm using it in this series.
- In Section 1.5, I discuss the probability that people will eventually make brain-like AGIs, as opposed to some other kind of AGI (or just not invent AGI at all). The section includes seven popular opinions on this topic, from both neuroscientists and AI / machine learning experts, and my responses.
- In Section 1.6, I'll address AGI accidents, which is something we should expect if we don't solve the AGI technical safety problem. I'll argue that these kinds of accidents can be catastrophic indeed, including human extinction. This topic is a *minefield* of confusion and miscommunication, and I will frame my discussion around responses to eight common objections.
- In Section 1.7, I'll address the more specific question of why we should think about AGI safety *right now.* After all, there is a *prima facie* good case for waiting, namely: (1) AGI doesn't exist yet, (2) AGI will exist someday in the future, and (3) it will be easier to do AGI safety research when we know more about the AGI, and easier still when we actually have AGI code that we can run tests on. There is indeed *something* to that argument, but I'll argue that there is nevertheless a lot of safety work that can and must be done ASAP.
- In Section 1.8, I'll suggest that brain-like-AGI safety is a fun, fascinating, and fruitful topic, even if you don't buy the idea that it's important for the future.

1.2 The AGI technical safety problem

AGI is short for "Artificial General Intelligence"—I'll get back to the definition of AGI in Section 1.4 below. AGI doesn't exist right now, but I'll argue in Section 1.7 that we can and should be preparing for AGI even today.

The part I'll be talking about in this series is the red box here:



Specifically, we zoom in on a single team of humans who are trying to create a single AGI, and we want it to be possible for them to do so without winding up with some catastrophe that nobody wanted, with an out-of-control AGI self-replicating around the internet or whatever (more on which in Section 1.6).

Blue boxes in this diagram are things that I won't talk about in this series. In fact, I'm not working on them at all—I have enough on my plate already. But I very strongly endorse other people working on them. If *you*, dear reader, want to work on them, *godspeed*!! I'm cheering you on! And here are a few links to get you started: 1, 2, 3, 4, 5, 6, 7.

Back to the red box. This is a technical problem, calling for a technical solution. Nobody wants catastrophic accidents. And yet!! Indeed, it's entirely possible for people to write an algorithm that does something that nobody wanted it to do. It happens all the time! **We might call it "a bug" when it's a local problem in the code, and we might call it "a fundamentally flawed software design" when it's a global problem.** I'll argue later in the series that AGI code may be unusually prone to catastrophic accidents, and that the stakes are very high (see Section 1.6 below, and Post #10).

Here's an analogy. If you're building a nuclear power plant, nobody wants an out-of-control chain reaction. The people at Chernobyl certainly didn't! But it happened anyway! I take a few lessons from this analogy:

- Enrico Fermi invented a technical solution for controlling nuclear chain reactions—control rods—*before* starting to build the first-ever nuclear chain reaction. Right on!! *That's* doing things in the right order! By the same token, I suggest that we should strive to have a technical solution to avoiding catastrophic AGI accidents ready to go *before* people start programming AGIs. In fact, I'll argue below for something even stronger than that: knowing the solution (even vaguely) 10 years before AGI is even better; 20 years before AGI is better still; etc. etc. This claim is not obvious, but I'll get back to it (Section 1.7).
- Technical solutions aren't all-or-nothing. Some reduce the chance of accidents without eliminating them. Some are complicated and expensive and error-prone to implement. In the nuclear case, control rods reduce accident risk a lot, but passively-safe reactors reduce it even further. By the same token, I expect that technical AGI safety will be a

rich field, where we'll develop better and better approaches over time, involving multiple techniques and multiple layers of protection. At least, I hope! As I'll discuss later in the series, I claim that right now we have no solution at all—not even vaguely. We have our work cut out!

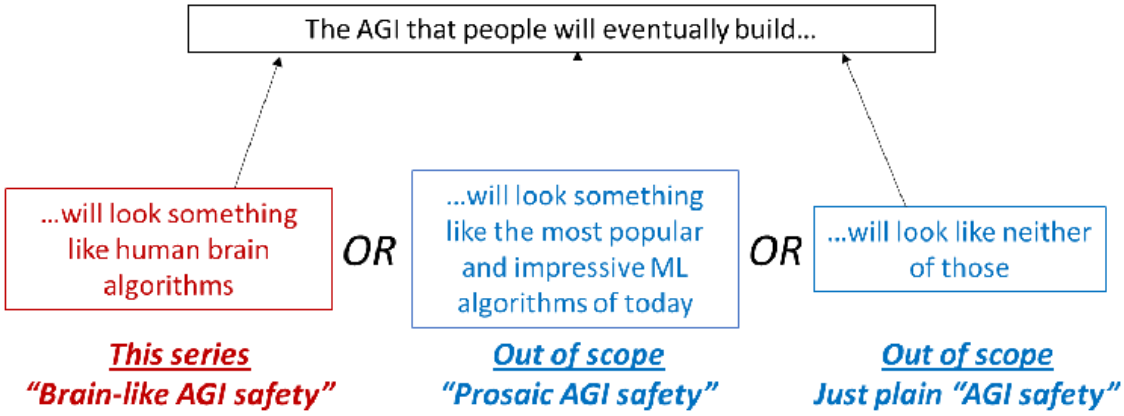
- The blue boxes (see diagram above) also exist, and are absolutely essential, even if they're out-of-scope for this particular series of articles. The cause of the Chernobyl accident was *not* that nobody knew how to keep a nuclear chain reaction under control, but rather because best practices were not followed. In that case, all bets are off! Still, although we on the technical side can't *solve* this noncompliance problem by ourselves, we can *help on the margin*, by developing best practices that are maximally idiot-proof, and minimally expensive.

In *The Sorcerer's Apprentice*, if I'm remembering it right, software engineer Mickey Mouse programs an AGI with a broomstick-like robotic body. The AGI proceeds to do the exact thing that Mickey *programmed* it to do ("fill a bucket with water"), but that turns out to be very different from what Mickey *wanted* it to do ("fill a bucket with water, without making a mess or doing anything else that I would find problematic, etc.") Our goal is to *empower* software engineers like Mickey with the technical tools and knowledge they need to avoid these kinds of accidents. See this talk by Nate Soares for a deep-dive into why Mickey has his work cut out.

1.3 Brain-like AGI

1.3.1 Overview

This series will focus on a particular scenario for what AGI algorithms will look like:



The red box is what I'll talk about here. The blue boxes are things that are out-of-scope for this series.

You may have opinions about which of these categories is more or less likely, or impossible, or whether this breakdown is even sensible. I have opinions about those things too! I'll discuss them later (Section 1.5). My main opinion is that *all three* of these are sufficiently likely that we should be "contingency planning" for them. So while I personally don't do too much work on the blue boxes, I'm sure glad that other people do!

Here's an analogy. If someone in 1870 were guessing what future human flight would look like…

- *"Kinda like birds"* would have been a reasonable guess…
- *"Kinda like today's best airships"* would *also* have been a reasonable guess…
- *"Neither of the above"* would have been a reasonable guess too!!

In this particular imaginary case, all three of those guesses would have turned out correct in some ways and wrong in other ways: The Wright Brothers were directly and extensively inspired by large soaring birds, but left out the wing-flapping part. They also used some components found on airships (e.g. propellers), as well as plenty of original ingredients. That's just one example, but I think it's suggestive.

1.3.2 What exactly is "brain-like AGI"?

When I say "brain-like AGI", I have a particular thing in mind. This thing will become much clearer in the subsequent posts, after we've started diving into neuroscience. But here's what I'm going for in broad strokes.

There are ingredients in the human brain and its environment that lead to humans having general intelligence (i.e., common sense, ability to figure things out, etc.—see Section 1.4 below). The scenario I have in mind is: Researchers will figure out what those ingredients in the brain are, and how they work, and then they will write AI code based on those same key ingredients.

To clarify:

- I don't expect that "brain-like AGI" will include *every* part of the brain and its environment. For example, there are highly-intelligent people who were born without a sense of smell, which suggests that brain olfactory processing circuitry probably isn't essential for AGI. There are highly-intelligent people who were quadriplegic from birth,

suggesting that lots of spinal-cord circuitry and (certain aspects of) "embodiment" aren't essential either. There are likewise people born without a cerebellum who are nevertheless well within the range of normal adult human intelligence (able to hold down a job, live independently, etc.—the kinds of capabilities that we would unhesitatingly call "AGI"). Other adults are holding down jobs while missing an entire brain hemisphere, etc. etc. My default expectation is that AGI will be created by people trying to create AGI, and they'll leave out whatever components they can, to make their jobs easier. (I'm not endorsing that as necessarily a good idea, just saying what I expect by default. More on this in Post #3.)
    - In particular, the kind of "brain-like AGI" I'm talking about is *definitely* not the same as Whole Brain Emulation.
- I don't require that "brain-like AGI" will resemble the human brain in low-level details, like with spiking neurons, dendrites, etc., or direct simulations thereof. If the resemblance is only at a higher level of abstraction, that's fine, it won't affect anything here.
- I don't require that "brain-like AGI" will be invented by a process of reverse-engineering the brain. If AI researchers happen to *independently* reinvent brain-like algorithms—just because those algorithms are good ideas—well then I still count it as brain-like algorithms.
- I don't require that "brain-like AGI" will be designed in a way that resembles how the brain was designed, i.e. evolutionary search. Quite the contrary: My working assumption is that it will be designed by humans in a way that's akin to a typical machine learning project today: lots of human-written code (loosely analogous to the genome), a *subset* of which defines the inference and update rules of one or more learning algorithms (corresponding to the brain's *within-lifetime* learning algorithms). There may be a few blank spaces in the code that get filled in by hyperparameter search or neural architecture search etc. Then you run the code, and the learning algorithms gradually build up a big complicated trained model from experience, maybe with trillions of adjustable parameters. Much more on this stuff in the next two posts and Post #8.
- I don't require that "brain-like AGI" will be conscious (in the phenomenal sense). There are *ethical* reasons to care about whether AGI is conscious (more on which in Post #12), but nothing I say in this series will depend on whether or not the AGI is conscious. Machine consciousness is a big contentious topic and I just don't want to get into it here. (I've written about it a bit elsewhere.)

Maybe a more practical way of saying it is: I'm going to make a bunch of claims about the algorithms underlying human intelligence, and then talk about safely using algorithms with those properties. If our future AGI algorithms have those properties, then this series will be useful, and I would be inclined to call such an algorithm "brain-like". We'll see exactly what those algorithm properties are, going forward.

1.3.3 "Brain-like AGI" (by my definition) can (and quite possibly will) have *radically nonhuman motivations*

I'm going to talk about this a *lot* more in later articles, but this is *such* an important point that I want to bring it up immediately.

Yes I know it sounds weird.

Yes I know you think I'm nuts.

But please, I beg you, hear me out first. By the time we get to Post #3, well *then* you can decide whether or not to believe me.

In fact, I'll go further. I'll argue that "radically nonhuman motivations" is not just *possible* for a brain-like AGI, but is *my baseline expectation* for a brain-like AGI. I'll argue that this is generally a bad thing, and that we should consider prioritizing certain lines of R&D in a proactive effort to avoid that.

(To be clear, "radically nonhuman motivations" is not synonymous with "scary and dangerous motivations". Unfortunately, "scary and dangerous motivations" is *also* my baseline expectation for a brain-like AGI!! But that requires a further argument, and you'll have to wait until Post #10 for that one.)

1.4 What exactly is "AGI"?

A frequent point of confusion is the word "General" in "Artificial General Intelligence":

- **The word "General" DOES mean "not specific"**, as in "In general, Boston is a nice place to live."
- **The word "General" DOES NOT mean "universal"**, as in "I have a general proof of the math theorem."

An AGI is *not* "general" in the latter sense. It is *not* a thing that can instantly find every pattern and solve every problem. Humans can't do that either! In fact, *no* algorithm can, because that's fundamentally impossible. Instead, an AGI is a thing that, when faced with a difficult problem, might be able to solve the problem easily, but if not, maybe it can build a tool to solve the problem, or it can find a clever way to avoid the problem altogether, etc. For our purposes here, think of AGI as an algorithm which can "figure things out" and "understand what's going on" and "get things done", including using language and science and technology, in a way that's reminiscent of how most adult humans can do those things, but toddlers and chimpanzees and GPT-3 can't. Of course, AGI algorithms may well be subhuman in some respects and superhuman in other respects.

Anyway, this series is about brain-like algorithms. These algorithms are by definition capable of doing absolutely every intelligent behavior that humans can do, and potentially much more. So they can *definitely* reach AGI. Whereas today's AI algorithms are *not* AGI. So somewhere in between here and there, there's a fuzzy line that separates "AGI" from "not AGI". Where exactly is that line? My answer: I don't know, and I don't care. Drawing that line has never come up for me as a useful thing to do. It won't come up in this series either.

1.5 What's the probability that we'll eventually wind up with brain-like AGI?

Above (Section 1.3.1) I suggested three categories of AGI algorithms: "brain-like" (as defined just above), "prosaic" (i.e. like today's most impressive deep neural net ML algorithms), and "other".

If our attitude is "Yes, let's do safety research for all three possibilities, just in case!!"—as I claim it should be—then I guess it's not all *that* decision-relevant what probability weights we put on each of the three things.

But even if it's irrelevant, it's fun to talk about, so what the heck, I'll just quickly summarize and respond to some popular opinions I've heard on this topic.

Opinion #1: *"I dispute the premise: human brains work by basically the same principles as today's popular ML algorithms."*

- The thing is, “today’s popular ML algorithms” is a big tent including lots of different algorithms. For example, I see *hardly any* overlap between "GPT-3-like-AGI safety" and "brain-like-AGI safety", whereas I see *substantial* overlap between "model-based RL AGI safety" and "brain-like-AGI safety".
- Anyway, by assuming “brain-like AGI”, I get the right to make certain assumptions about the cognitive architecture, representations, learning algorithms, and so on.
    - *Some* of these “brain-like AGI ingredients” are *universal* parts of today’s popular ML algorithms (e.g. learning algorithms; distributed representations).
    - *Others* of these “brain-like AGI ingredients” are (individually) present in a *subset* of today’s popular ML algorithms but absent from others (e.g. reinforcement learning; predictive [a.k.a. self-supervised] learning; explicit planning).
    - *Still others* of these “brain-like AGI ingredients” seem mostly or totally absent from today’s most popular ML algorithms (e.g. ability to form "thoughts" [e.g. "I’m going to the store"] that blend together immediate actions, short-term predictions, long-term predictions, and flexible hierarchical plans, inside a generative world-model that supports causal and counterfactual and metacognitive reasoning).
- So in this sense, “brain-like AGI” is a specific thing that might or might not happen, independently of “prosaic AGI”. Much more on “brain-like AGI”, or at least its safety-relevant aspects, in the subsequent posts.

Opinion #2: *“Brain-like AGI is possible but Prosaic AGI is not. It just ain’t gonna happen. Today's ML research is not a path to AGI, just as climbing a tree is not a path to the moon.”*

- I find this to be a moderately popular opinion among neuroscientists and cognitive psychologists. Prominent advocates of this view include Gary Marcus and Melanie Mitchell.
- One question is: if we take one of today’s most popular ML models, add no additional significant insights or architectural changes whatsoever, and just scale the model to ever larger sizes, do we get AGI? I join those neuroscientists in expecting the answer to be “probably not”.
- On the other hand, *even if* it turns out that deep neural networks can't do important-for-intelligence things X and Y and Z, well c’mon, somebody's probably just gonna glue together a deep neural network with other components that do X and Y and Z. And then we can have some pointless semantic debate about whether it's still “really” prosaic AGI or not.

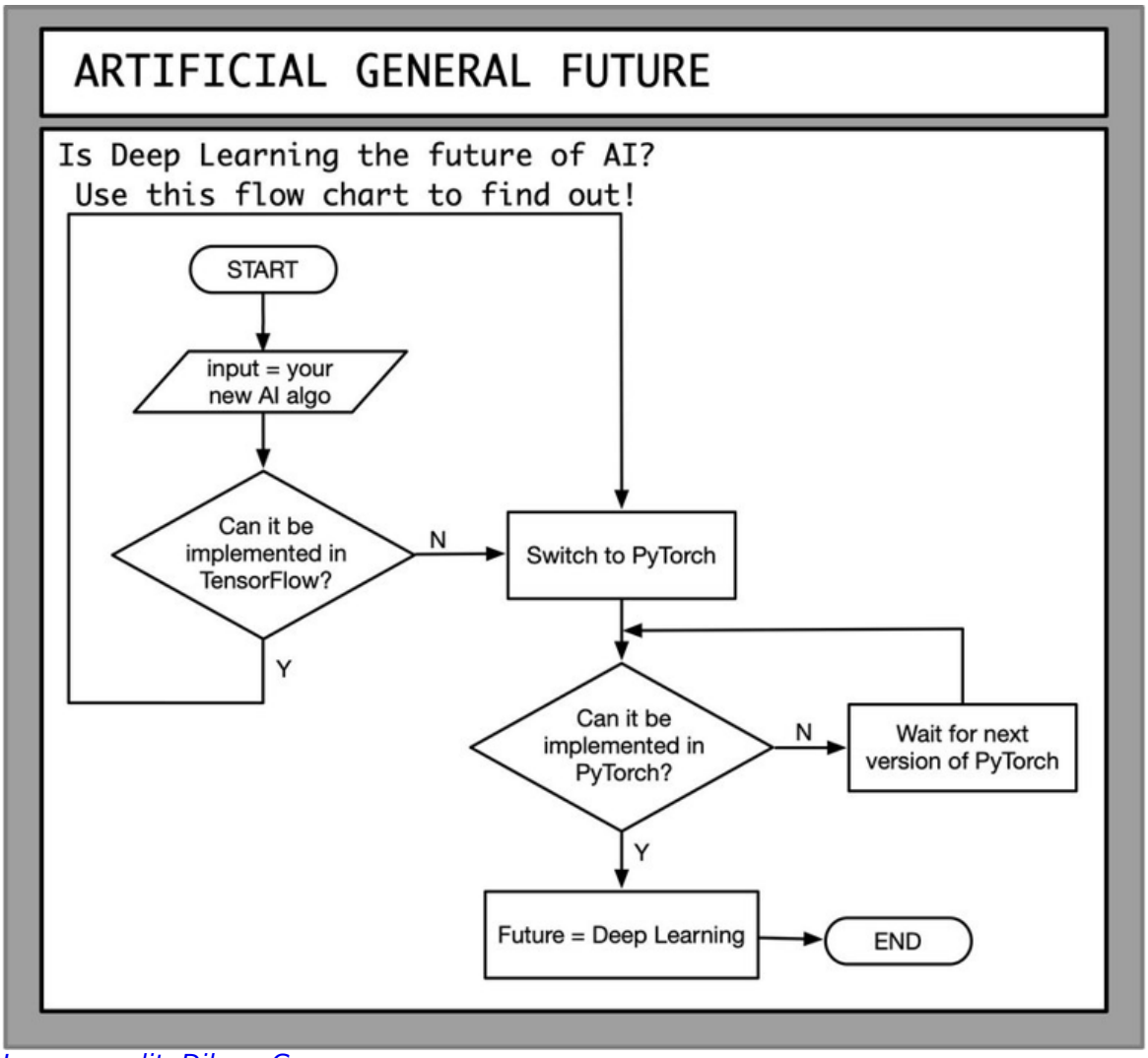


Image credit: Dileep George

- Anyway, in this series, I will be assuming that AGI will have certain algorithmic features (e.g. online learning, a certain type of model-based planning, etc.—much more in later posts). I'll be assuming that because (1) those features are part of human intelligence, (2) they seem to be there for a good reason. My safety-related discussions will rely on those features being present. Can algorithms with these features be implemented in PyTorch on a GPU? Well, I don't really care.

Opinion #3: *"Prosaic AGI is going to happen so soon that no other research program has a chance."*

- A subset of people in ML believe this. I don't. Or at any rate, I would be awfully surprised.
- I do agree that *IF* prosaic AGI is, say, 5 years away, then we almost certainly don't need to think about brain-like AGI or indeed any other research program. I just think that's an awfully big "if".

Opinion #4: *"Brains are SO complicated—and we understand them SO little after SO much effort—that there's just no way we'll get brain-like AGI even in the next 100 years."*

- This is a pretty popular opinion, both inside and outside of neuroscience. I think it's very wrong, and will be arguing against it at length in the next two posts of the series.

Opinion #5: *"Neuroscientists aren't trying to invent AGI, so we shouldn't expect them to succeed."*

- There's *some* truth to this, but I mostly disagree. For one thing, a number of leading computational neuroscientists (the DeepMind neuroscience team, Randall O'Reilly, Jeff Hawkins, Dileep George) *are* in fact explicitly trying to invent AGI. For another thing, people *in AI*, including prominent leaders of the field, try to keep up with the neuroscience literature and incorporate its ideas. And anyway, "understanding an AGI-relevant brain algorithm" is *part* of inventing brain-like AGI, whether or not that's the intention of the person carrying out the research.

Opinion #6: *"Brain-like AGI is kinda an incoherent concept; intelligence requires embodiment, not just a brain in a vat (or on a chip)."*

- The "embodiment" debate in neuroscience continues to rage. I fall somewhere in the middle. I do think that future AGIs will have *some* action space—e.g., the ability to (virtually) summon a particular book and open it to a particular passage. I *don't* think having a whole literal body is important—for example Christopher Nolan (1965-2009) had lifelong quadriplegia, but it didn't prevent him from being an acclaimed author and poet. More importantly, I expect that whatever aspects of embodiment are important for intelligence could be easily incorporated into a brain-like AGI running on a silicon chip. Is a body necessary for intelligence after all? OK sure, we can give the AGI a virtual body in a VR world. Are hormonal signals necessary for intelligence? OK sure, we can code up some virtual hormonal signals. Etc. etc.

Opinion #7: *"Brain-like AGI is incompatible with conventional silicon chips; it requires a whole new hardware platform based on spiking neurons, active dendrites, etc. Neurons are just plain better at computation than silicon chips are—just look at their energy efficiency etc."*

- I'm really unsympathetic to this position. Conventional silicon chips can definitely simulate biological neurons—neuroscientists do this all the time. Conventional silicon chips can also presumably implement "brain-like algorithms" using different low-level operations more suited to that hardware, just as the same C code can be compiled to different CPU instruction sets. As for "neurons are just plain better", I freely acknowledge the human brain does a *crazy impressive* amount of computation given its tiny volume, mass, and power consumption. But those are not hard constraints! If a silicon-chip AGI server were literally 10,000× the volume, 10,000× the mass, and 1000× the power consumption of a human brain, with comparable performance, I don't think anyone would be particularly bothered—in particular, its electricity costs would *still* be well below my local minimum wage!! And my best estimate is that buying enough silicon chips for human-brain-human-lifetime-level computation is probably easily feasible, or will be in the next decade, even for small companies. The key reason that small companies aren't building AGIs today is that we don't know the right algorithms.

This is just a quick run-through; each of these opinions could be a whole article—heck, a whole book. For my part, I put >50% probability that we'll have a *sufficiently*-brain-like AGI that this series will be very relevant. But who knows, really.

1.6 Why are AGI accidents such a big deal?

Two reasons: (1) the stakes are high, and (2) the problem is hard. I'll be talking about (2) much more later in the series (Posts #10–#11). Let's talk about (1).

And let's talk more specifically about one high-stakes possibility: the risk of human extinction. That sounds a bit wild but hear me out.

I'll frame this discussion as answers to popular objections:

Objection #1: **The only way that an out-of-control AGI could result in human extinction is if the AGI invents crazy sci-fi superweapons, e.g. gray goo . As if such a thing is even possible!**

Oh, if only that were true! But alas, I don't think sci-fi superweapons are necessary. In fact, it seems to me that it's maybe borderline possible for a *human* intelligence using *existing* technology to cause human extinction!

Think about it: it's already at least borderline-possible today for an ambitious intelligent charismatic methodical human to arrange for the manufacture and release of a novel contagious disease that's 100× deadlier than COVID-19. Heck, it's probably possible to release 30 such plagues all at once! Meanwhile, I figure it's at least borderline-possible today for an ambitious intelligent charismatic methodical human to find a way to manipulate nuclear early warning systems (trick them, hack into them, bribe or threaten their operators, etc.), setting off an all-out nuclear war, killing billions of people and sowing chaos around the world. Those are just two things; creative readers will immediately think of lots more. I mean seriously, there are fiction books with *totally plausible* mad-scientist apocalypse scenarios—not just according to me, but according to domain experts.

Now, granted, human extinction seems like a high bar! People live in all kinds of places, including small tropical islands that would be insulated from both nuclear winter and plagues. But this is where we get a big difference between an *intelligent* agent like an AGI, versus an *unintelligent* agent like a virus. Both can self-replicate. Both can kill lots of people. But an AGI, unlike a virus, can *take command of military drones, and mow down the survivors!!*

So my hunch is that we're all still around today thanks in large part to the fact that all the most ambitious intelligent charismatic methodical humans *aren't trying to kill everyone*—and not because "killing everyone" is a thing that requires crazy sci-fi superweapons.

As discussed above, one of the failure modes I have in mind would involve out-of-control AGIs that combine (at least) human-like intelligence with *radically nonhuman motivations*. This would be a new situation for the world, and I don't find it comforting!

You might reply: The thing that went wrong in this scenario is *not* the out-of-control AGI, it's the fact that humanity is too vulnerable! And my response is: *Why can't it be both?* So in my book: Yes we should absolutely make humanity more robust to bio-engineered pandemics, and reduce the chances of nuclear war, etc. etc. All these things are great ideas that I strongly endorse, and *godspeed* if you yourself are working on them. But at the same time, we should *also* work really hard to *not create out-of-control self-replicating human-like intelligences with radically nonhuman motivations*!

…Oh and one more thing: Maybe "crazy sci-fi superweapons like gray goo" are possible too! Beats me! If so, we need to be *even more* cautious!

Objection #2: **The only way that an AGI accident could result in human extinction is if the AGI is somehow smarter than *all humans combined* .**

The issue here is that "all humans combined" may not know that they are engaged in a battle against an AGI. Maybe they would, maybe they wouldn't. If the AGI is at all competent at secrecy, it could presumably set up a surprise attack, without anyone knowing what was happening until it was too late. Or if the AGI is at all competent at disinformation and propaganda, it could presumably pass off its actions as accidents, or as (human) enemy

action. Maybe everybody would be blaming everyone else, and nobody would know what's going on.

Objection #3: **The only way that an AGI accident could result in human extinction is if the AGI is deliberately given access to levers of power, like nuclear codes, control over social media, etc. By the same token, we can run the AGI code on just one server, and then switch it off if anything goes wrong.**

The problem here is that intelligent agents can turn "few resources" into "lots of resources". Think of Warren Buffett, or Adolf Hitler.

Intelligent agents can earn money (whether legally or not), and they can earn trust (whether deserved or not), and they can get access to other computers (whether by purchasing server time or by hacking). The latter is especially important because an AGI—like a virus, but *not* like a human—can potentially self-replicate. Self-replication is one way it can protect itself from shutdown, if it's motivated to do so. Another way is by tricking / misleading / winning over / bribing / outsmarting whoever controls the shutdown switch.

(A kernel of truth here is that if we're unsure of an AGI's motivations and competence, then *giving it access to the nuclear codes is a very bad idea*! Trying to limit an AGI's power and resources doesn't seem to be a solution to any of the *hardest* problems that we're interested in here, but it can still be helpful on the margin, like as an "additional layer of protection". So I'm all for it.)

Objection #4: **The good AGIs can stop the bad out-of-control AGIs.**

For one thing, if we don't solve the technical problem of how to steer an AGI's motivation and keep it under control (see Posts #10–#15), then there may be a period of time when *there are no good AGIs*! Instead, *all* the AGIs are out-of-control!

For another thing, out-of-control AGIs will have asymmetric advantages over good AGIs—like the ability to steal resources, to manipulate people and institutions via lying and disinformation; to cause wars, pandemics, blackouts, famines, gray goo, and so on; and to not have to deal with coordination challenges across different (human) actors with different beliefs and goals. More on this topic here.

Objection #5: **An AGI that's trying to kill everyone is a *really specific* kind of failure mode! There's just no reason that an AGI would try to do that. It's *not* the kind of thing that would happen as a general result of buggy or poorly-designed AGI software. It's the kind of thing that would *only* happen if somebody went out of their way to put malign motivations into the AGI. As a matter of fact, buggy or poorly-designed software tends to do, well, nothing in particular! I happen to know a thing or two about buggy software—in fact I just created some this morning. The only thing it murdered was my self-confidence!**

A kernel of truth here is that *some* bugs or design flaws in AGI code will indeed manifest as software that is *not* an AGI, and not "intelligent", and probably not even functional! Such errors do not qualify as catastrophic accidents, unless we were foolish enough to put that software in charge of the nuclear arsenal. (See "Objection #3" above.)

However, I claim that *other* bugs / design errors *will* in fact possibly lead to the AGI deliberately killing everyone, even if the AGI designers are reasonable people with noble, humble intentions.

Why? In the AGI safety lore, the classic way to justify this claim is the trifecta of (1) "The Orthogonality Thesis", (2) "Goodhart's law", and (3) "Instrumental Convergence". You can get the short version of this three-part argument in this talk. For the long version, read on: this series is all about the nuts and bolts of motivation in brain-like AGIs, and how it can go awry.

So, hold that thought, and all will be clear by the time we get through Post #10.

Objection #6: **If building AGIs seems to be a catastrophic-accident-prone endeavor, we'll just stop doing it, until when (and if) the problem is solved.**

My immediate reaction is to say: “We”? Who the heck is "*we*"? The AI community consists of many thousands of skilled researchers scattered across the globe. They disagree with each other about practically everything. There is no oversight on what they’re doing. Some of them work at secret military labs. So I don’t think we can take it for granted that “we” will not engage in research that *you and I* consider to be obviously ill-conceived and risky.

(Also, if some catastrophic accidents can be unrecoverable, then even one of those is too many.)

By the way, suppose someone says to me: “I have an extraordinarily ambitious plan, one that will require many years or decades of groundwork, but *if we succeed* then “Everyone on earth pauses AGI R&D until safety problems are resolved” will be on the table as a possible policy option in the future.” OK sure, I would listen to that person with an open mind. They seem like they’re at least understanding the scale of the challenge. Of course, I would expect them to probably fail. But what do I know?

Objection #7: **Accident risks have been going down and down, for decades. Didn’t you read Steven Pinker ? Have faith!**

Accident risks don’t solve themselves. They get solved when people solve them. Planes generally don’t crash because people have figured out how to avoid plane crashes. Nuclear power plants generally don’t melt down because people have figured out how to avoid nuclear meltdowns.

Imagine if I said, “Good news, car accident death rates are lower than ever! So now we can get rid of seatbelts and crumple zones and road signs!” You would respond: “No!! That’s insane!! Seatbelts and crumple zones and road signs *are the very reason* that car accident death rates are lower than ever!”

By the same token, if you’re optimistic that we’ll ultimately avoid AGI accidents, that’s not a reason for you to be opposed to AGI safety research.

There’s another thing to keep in mind before you take comfort in the historical record on technological accident risk: as technology gets inexorably more powerful, the scope of damage from technological accidents gets inexorably bigger as well. A nuclear bomb accident would be worse than a conventional bomb accident. A bioterrorist using 2022 technology would be able to do far more damage than a bioterrorist using 1980 technology. So by the same token, as AI systems get dramatically more powerful in the future, we should expect the scope of damage from AI accidents to grow dramatically as well. Thus the historical record here is not necessarily indicative of the future.

Objection #8: **Humans are doomed anyway. Oh well, whatever, no species lasts forever.**

I hear variants on this a lot. And granted, I can’t *prove* that it’s wrong. But the horseshoe crab has been around almost half a billion years, and counting. C’mon people, we can do this! Well at any rate, *I’m* not going down without a fight!

As for the people taking a “far mode” detached armchair-philosopher attitude to human extinction: If you would be devastated by the untimely death of your best friend or beloved family member … but you’re not particularly bothered by the idea of an out-of-control AGI killing everybody … … umm, I’m not sure what to say here. Maybe you’re not thinking things through very carefully?

1.7 Why think about AGI safety *now*? Why not wait until we're closer to AGI and hence know more?

This is a common objection, and it indeed has a giant kernel of truth: Namely, that in the future, when we know more details about the eventual AGI design, there will be a lot of *new* technical safety work to do—work that we can't do right now.

However, there *is* safety work we can do right now. Just keep reading this series if you don't believe me!

I want to argue that the safety work that we *can* do right now, we *really should* do right now. Waiting would be much worse—even if AGI is still many decades away. Why's that? Three reasons:

Reason 1 for feeling a sense of urgency: *Early hints about safety can inform early R&D decisions—including via "Differential Technological Development".*

The most important thing is that there's certainly more than one way to code an AGI algorithm.

Very early on in the process, we're making decisions about the big-picture path to AGI. We could do R&D towards one of many variations on "brain-like AGI" as defined here, versus whole brain emulation, versus various types of "prosaic AGI" (Section 1.3.1), versus graph database query something-or-other AGI, versus knowledge / discussion / reasoning systems, and we can proceed with or without brain-computer interfaces of various types, and so on. Probably not *all* of these research paths are feasible, but there's certainly more than one path towards more than one possible destination. We get to pick which one to go down.

Heck, we get to decide whether to build AGI in the first place! (However, see "Objection #6" above.)

In fact, we're making these decisions already today. We've been making them for years. And our decision procedure is that lots of individuals around the world ask: What R&D direction is best for *me* right *now*? What gets *me* a job / promotion / profit / high-impact journal publication right *now*?

A *better* decision procedure would be: What kind of AGI do we eventually want to build? OK! Let's try to make *that* one happen, sooner than all the inferior alternatives.

In other words, if someone chooses an R&D direction based on whatever looks interesting and promising, just like everyone else, well they're not going to change our eventual technology development path. They're just going to move us down the *same* path slightly *faster*. If we think that some destinations are better than others—say, if we're trying to avoid a future full of out-of-control AGIs with radically nonhuman motivations—then it's important to pick and choose what research you're doing, in order to strategically accelerate the things that we most want to happen. This principle is called **differential technological development**—or more generally, **differential intellectual progress**.

I have my own preliminary ideas about what should be accelerated for brain-like AGI to go better. (I'll get to it much later in the series.) But **the *main* thing I believe is: "We should differentially accelerate work towards figuring out which work should be differentially accelerated"!! For example, would brain-like AGI be catastrophic-accident-prone or not? We have to figure it out! Hence this series!**

Reason 2 for feeling a sense of urgency: *We don't know how long safety research will take.*

As discussed much more in later posts (especially Posts #10–#15), it is currently unknown how to make an AGI which is reliably trying to do the things that we want it to be trying to do. We don't know how long it will take to figure it out (or prove that it's impossible!). It seems prudent to start now.

As discussed later in the series (especially Posts #10–#15), AGI Safety seems to be a gnarly technical problem. We don't currently know how to solve it—in fact, we don't even

> know if it's solvable. Thus, it seems wise to sharpen our pencils and get to work right now, rather than waiting until the last possible second. *[Meme concept stolen from here.]*

In Stuart Russell's memorable analogy, imagine that we get a message from the aliens: *"We are coming in our spaceships, and will arrive in 50 years. When we get there, we will radically transform your whole world beyond recognition."* Indeed, we see their ships in our telescopes. They're inching closer each year. What do we do?

If we were to respond to the coming alien invasion the way we are *actually* today responding to AGI, we would collectively shrug and say "Meh, 50 years, I mean, that's *really* far away. We don't have to think about that *now*! If 100 people on Earth are trying to prepare for the looming alien invasion, that's *plenty*. Maybe too much! Y'know, if you ask me, those 100 people on Earth should stop looking up at the stars, and look around their own communities. Then they'd see that the *REAL* 'looming alien invasion' is *cardiovascular disease*. That's killing people *right now*!"

…You get the idea. (Not that I'm bitter or anything.)

Reason 3 for feeling a sense of urgency: *Building near-universal consensus about anything can be a horrifically slow process.*

Suppose I have a really good and correct argument that some AGI architecture or approach is just a terrible idea—that it's unfixably unsafe. I publish the argument. Will everyone involved in AGI development, including those who have invested their career in that approach, immediately believe me, and change course? Probably not!!

That kind of thing *does* happen sometimes, especially in mature fields like math. But other ideas take many decades to become widely (let alone universally) accepted—famous examples include evolution and plate tectonics. It takes time for arguments to be refined. It takes time for evidence to be marshaled. It takes time for nice new pedagogical textbooks to be created. And yes, it takes time for the stubborn holdouts to die and be replaced by the next generation.

Why is near-universal consensus so important? See Section 1.2 above. Good ideas about how to build AGI are pointless if the people building AGI don't follow them. If we're going for voluntary compliance, then we need the AGI-builders to believe the ideas. If we're going for mandatory compliance, then we need the people with political power to believe the ideas. And we would *still* need AGI-builders to believe the ideas too, because perfect enforcement is a pipe dream (especially given secret labs etc.).

1.8 …Plus it's a really fascinating problem!

Hey neuroscientists, listen. Some of you are trying to cure diseases. Good for you. Have at it. Others of you, well, you *say* you're trying to cure diseases on your NIH grant applications, but c'mon, that's not your *real* goal, and everyone knows it. You're *really* in it to solve fascinating unsolved problems. Well, let me tell you, brain-like-AGI safety is a fascinating unsolved problem!

It's even a rich source of insights *about neuroscience*! When I'm thinking all day about AGI safety stuff (wireheading, wishful thinking, symbol-grounding, ontological crises, interpretability, blah blah blah), I'm asking very different questions than most

neuroscientists, and thus finding different ideas. (…I'd like to think. Well, read on, and you can decide for yourself whether they're any good.)

So even if I haven't convinced you that the technical AGI safety problem is super duper important and impactful, read on anyway. You can also work on the problem because it's awesome. ;-)

[Intro to brain-like-AGI safety] 2. “Learning from scratch” in the brain

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

2.1 Post summary / Table of contents

Part of the “Intro to brain-like-AGI safety” post series .

Having introduced the “brain-like-AGI safety” problem in the previous post, the next 6 posts (#2–#7) are primarily about neuroscience, building up to a more nuts-and-bolts understanding of what a brain-like AGI might look like (at least, in its safety-relevant aspects).

This post focuses on a concept that I call “learning from scratch”, and I will offer a hypothesized breakdown in which 96% of the human brain (including the neocortex) “learns from scratch”, and the other 4% (including the brainstem) doesn’t. This hypothesis is central to how I think the brain works, and hence will be a key prerequisite for the whole rest of the series.

- In Section 2.2, I’ll define the concept of “learning from scratch”. As an example, if I assert that the neocortex “learns from scratch”, then I’m claiming that the neocortex starts out totally useless to the organism—outputting fitness-improving signals no more often than chance—until it starts learning things (within the individual's lifetime). Here are a couple everyday examples of things that “learn from scratch”:
  - In most deep learning papers, the trained model “learns from scratch”—the model is initialized from random weights, and hence the model outputs are random garbage *at first*. But during training, the weights are updated, and the model outputs eventually become very useful.
  - A blank hard disk drive also “learns from scratch”—you can't pull useful information *out* of it until after you’ve written information *into* it.
- In Section 2.3, I will clarify some frequent confusions:
  - “Learning from scratch” is different from “blank slate”, because there is an innate learning algorithm, innate neural architecture, innate hyperparameters, etc.
  - “Learning from scratch” is different from “nurture-not-nature”, because (1) only *some* parts of the brain learn from scratch, while other parts don’t, and (2) the learning algorithms are not necessarily learning about the external environment—they could also be learning e.g. how to control one’s own body.
  - “Learning from scratch” is different from (and more specific than) “brain plasticity”, because the latter can also include (for example) a genetically-hardwired circuit with just one specific adjustable parameter, and that parameter changes semi-permanently under specific conditions.
- In Section 2.4, I’ll propose my hypothesis that two major parts of the brain exist solely to run learning-from-scratch algorithms—namely, the telencephalon (neocortex, hippocampus, amygdala, most of the basal ganglia, etc.) and cerebellum. Together these comprise 96% of the volume of the human brain.
- In Section 2.5, I’ll touch on four different lines of evidence concerning my hypothesis that the telencephalon and cerebellum learn from scratch: (1) big-picture thinking about how the brain works, (2) neonatal data, (3) a connection to the hypothesis of “cortical uniformity” and related issues, and (4) the possibility that a certain brain preprocessing motif—so-called “pattern separation”—involves randomization in a way that *forces* downstream algorithms to learn from scratch.
- In Section 2.6, I’ll talk briefly about whether my hypothesis is mainstream vs idiosyncratic. (Answer: I’m not really sure.)

- In Section 2.7, I'll offer a little teaser of why learning-from-scratch is important for AGI safety—we wind up with a situation where the thing we want the AGI to be trying to do (e.g. cure Alzheimer's) is a concept buried inside a big hard-to-interpret learned-from-scratch data structure. Thus, it is not straightforward for the programmer to write motivation-related code that refers to this concept. Much more on this topic in future posts.
- Section 2.8 will be the first of three parts of my *"timelines to brain-like AGI"* discussion, focusing on how long it will take for future scientists to reverse-engineer the key operating principles of the learning-from-scratch part of the brain. (The remainder of the timelines discussion is in the next post.)

2.2 What is "learning from scratch"?

As in the intro above, I'm going to suggest that large parts of the brain—basically the telencephalon and cerebellum (see Section 2.4 below)—"learn from scratch", in the sense that they start out emitting signals that are random garbage, not contributing to evolutionarily-adaptive behaviors, but over time become more and more immediately useful thanks to a within-lifetime learning algorithm.

Here are two ways to think about the learning-from-scratch hypothesis:

- **How you should think about learning-from-scratch (if you're an ML reader):** Think of a deep neural net initialized with random weights. Its neural architecture might be simple or might be incredibly complicated; it doesn't matter. And it certainly has an inductive bias that makes it learn certain types of patterns more easily than other types of patterns. But it still has to learn them! If its weights are initially random, then it's initially useless, and gets gradually more useful with training data. The idea here is that these parts of the brain (neocortex etc.) are likewise "initialized from random weights", or something equivalent.
- **How you should think about learning-from-scratch (if you're a neuroscience reader):** Think of a memory-related system, like the hippocampus. The ability to form memories is a very helpful ability for an organism to have! *…But it ain't helpful at birth!!*[1] You need to accumulate memories before you can use them! My proposal is that everything in the telencephalon and cerebellum are in the same category—they're kinds of memory modules. They may be *very special* kinds of memory modules! The neocortex, for example, can learn and remember a super-complex web of interconnected patterns, and comes with powerful querying features, and can even query itself in recurrent loops, and so on. But still, it's a form of memory, and hence starts out useless, and gets progressively more useful to the organism as it accumulates learned content.

2.3 Three things that "learning from scratch" is NOT

2.3.1 Learning-from-scratch is NOT "blank slate"

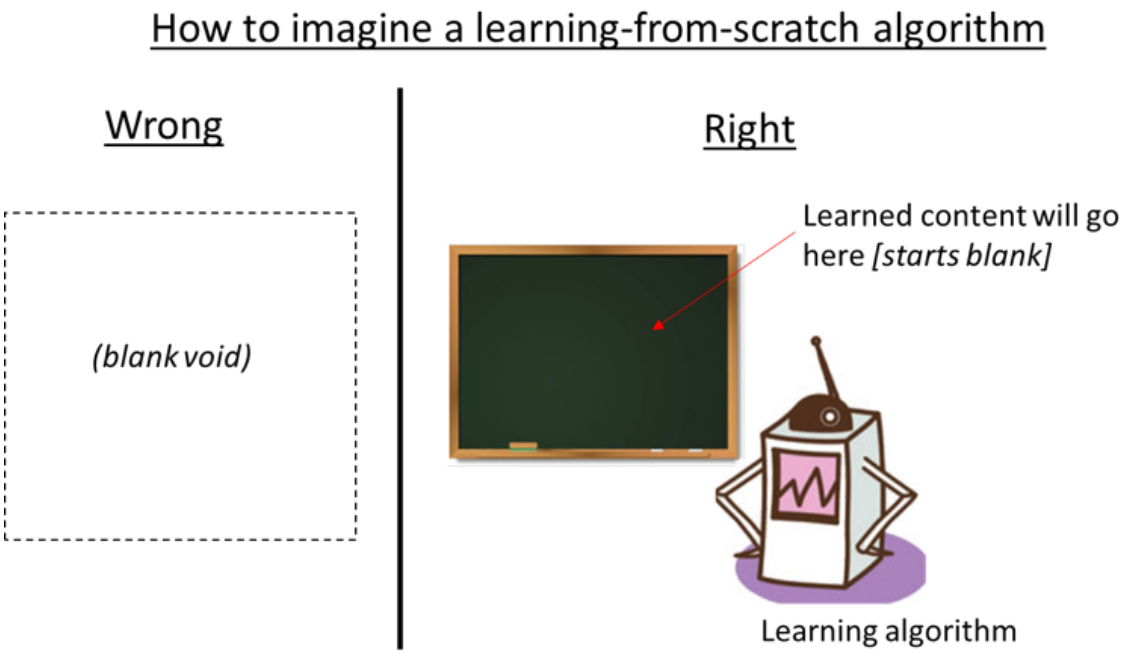
I already mentioned this, but I want to be crystal clear: if the neocortex (for example) learns from scratch, that does *not* mean that there is no genetically-hardcoded information content in the neocortex. It means that the genetically-hardcoded information content is probably better thought of as the following:

- *Learning algorithm(s)*—i.e., innate rules for semi-permanently changing the neurons or their connections, in a situation-dependent way.

- *Inference algorithm(s)*—i.e., innate rules for what output signals should be sent *right now*, to help the animal survive and thrive. The actual output signals, of course, will also depend on previously-learned information.
- *Neural network architecture*—i.e., an innate large-scale wiring diagram specifying how different parts of the learning module are connected to each other, and to input and output signals.
- *Hyperparameters*—e.g., different parts of the architecture might innately have different learning rates. These hyperparameters can also change during development (cf. "sensitive periods"). There can also be an innate capacity to change hyperparameters on a moment-by-moment basis in response to special command signals (in the form of neuromodulators like acetylcholine).

Given all those innate ingredients, the learning-from-scratch algorithm is ready to receive input data and supervisory signals from elsewhere[2], and it gradually learns to do useful things.

This innate information is not necessarily simple. There could be 50,000 wildly different learning algorithms in 50,000 different parts of the neocortex, and that would *still* qualify as "learning-from-scratch" in my book! (I don't think that's the case though—see Section 2.5.3 on "uniformity".)



When you imagine a learning-from-scratch algorithm, you should *not* imagine an empty void that gets filled with data. You should imagine an *automaton* that continually (1) writes information into a memory bank, and (2) performs queries on the current contents of the memory bank. "From scratch" just means that the memory bank starts out empty. There are *many* such automatons, each following a different procedure for exactly what to write and how to query. For example, a "lookup table" corresponds to a simple automaton that just records whatever it sees. Other automatons correspond to supervised learning algorithms, and reinforcement learning algorithms, and autoencoders, etc. etc.

2.3.2 Learning-from-scratch is NOT "nurture-over-nature"

There's a tendency to associate "learning-from-scratch algorithms" with the "nurture" side of the "nature-vs-nurture" debate. I think that's wrong. Quite the contrary: I think that the learning-from-scratch hypothesis is fully compatible with the possibility that evolved innate behaviors play a big role.

Two reasons:

First, *some* parts of the brain are *absolutely NOT* running learning-from-scratch algorithms! In this category is mainly the brainstem and hypothalamus (more about those below and in the next post). These non-learning-from-scratch parts of the brain would have to be fully responsible for any adaptive behavior at birth.[1] Is that plausible? I think so, given the impressive range of functionality in the brainstem. For example, the neocortex has circuitry for processing visual and other sensory data—but so does the brainstem! The neocortex has motor-control circuitry—but so does the brainstem! In at least some cases, full adaptive behaviors seem to be implemented entirely within the brainstem: for example, mice have a brainstem incoming-bird-detecting circuit wired directly to a brainstem running-away circuit. So my learning-from-scratch hypothesis is not making any *blanket* claims about what algorithms or functionalities are or aren't present in the brain. It's just a claim that certain types of algorithms are only in certain *parts* of the brain.

Second, "learning from scratch" is not the same as "learning from the environment". Here's a made-up example[3]. Imagine a bird brainstem is built with an innate capability to *judge* what a good birdsong should *sound* like, but lacks a *recipe* for how to *produce* a good birdsong. Well, a learning-from-scratch algorithm could fill in that gap—doing trial-and-error to get from the former to the latter. This example shows that **learning-from-scratch algorithms can be in charge of behaviors that we would naturally and correctly describe as innate / "nature not nurture"**.

2.3.3 Learning-from-scratch is NOT the more general notion of "plasticity"

"Plasticity" is a term for the brain semi-permanently changing itself, typically by changing the presence / absence / strength of neuron-to-neuron synapses, but also sometimes via other mechanisms, like changes of a neuron's gene expression.

Any learning-from-scratch algorithm necessarily involves plasticity. But not all brain plasticity is part of a learning-from-scratch algorithm. A second possibility is what I call "individual innate adjustable parameters". Here's a table with both an example of each and general ways in which they differ:

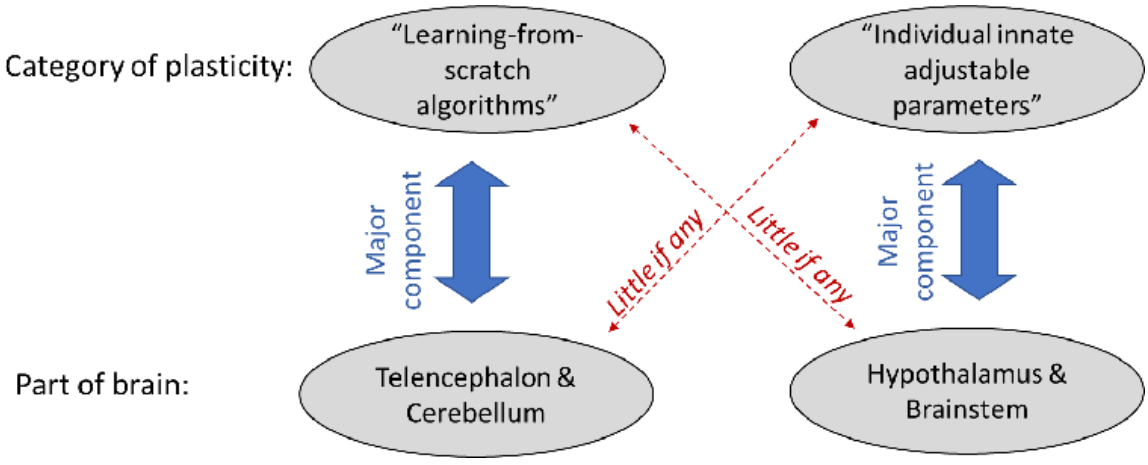
	Learning-from-scratch algorithms	**Individual innate adjustable parameters**
Stereotypical example to keep in mind:	Every deep learning paper: there's a *learning algorithm* that gradually builds a *trained model* by adjusting lots of parameters.	Some connection in the rat brain that strengthens when the rat wins a fight—basically, it's a tally of how many fights the rat has won over the course of its life. Then this connection is used to implement the behavior "If you've won lots of fights in your life, be more aggressive." (ref)
Number of	Maybe lots—	Probably few—even as few as one

parameters that change based on input data (i.e. how many dimensions is the space of all possible trained models?)	hundreds, thousands, millions, etc.	
If you could scale it up, would it work better after training?	Yeah, probably.	Huh?? WTF does “scale it up” mean?

I don’t think there’s a sharp line between these things; I think there’s a gray area where one blends into the other. Well, at least I think there’s a gray area *in principle*. In practice, I feel like it's a pretty clean division—whenever I learn about a particular example of brain plasticity, it winds up being clearly in one category or the other.

My categorization here, by the way, is a bit unusual in neuroscience, I think. Neuroscientists more often focus on low-level implementation details: “Does the plasticity come from long-term synaptic change, or does it come from long-term gene expression change?” “What’s the biochemical mechanism?” Etc. That’s a totally different topic. For example, I’d bet that the exact same low-level biochemical synaptic plasticity mechanism can be involved in both a learning-from-scratch algorithm and an individual innate adjustable parameter.

Why do I bring this up? Because I’m planning to argue that the hypothalamus and brainstem have little or no learning-from-scratch algorithms, so far as I can tell. But they *definitely* have individual innate adjustable parameters.



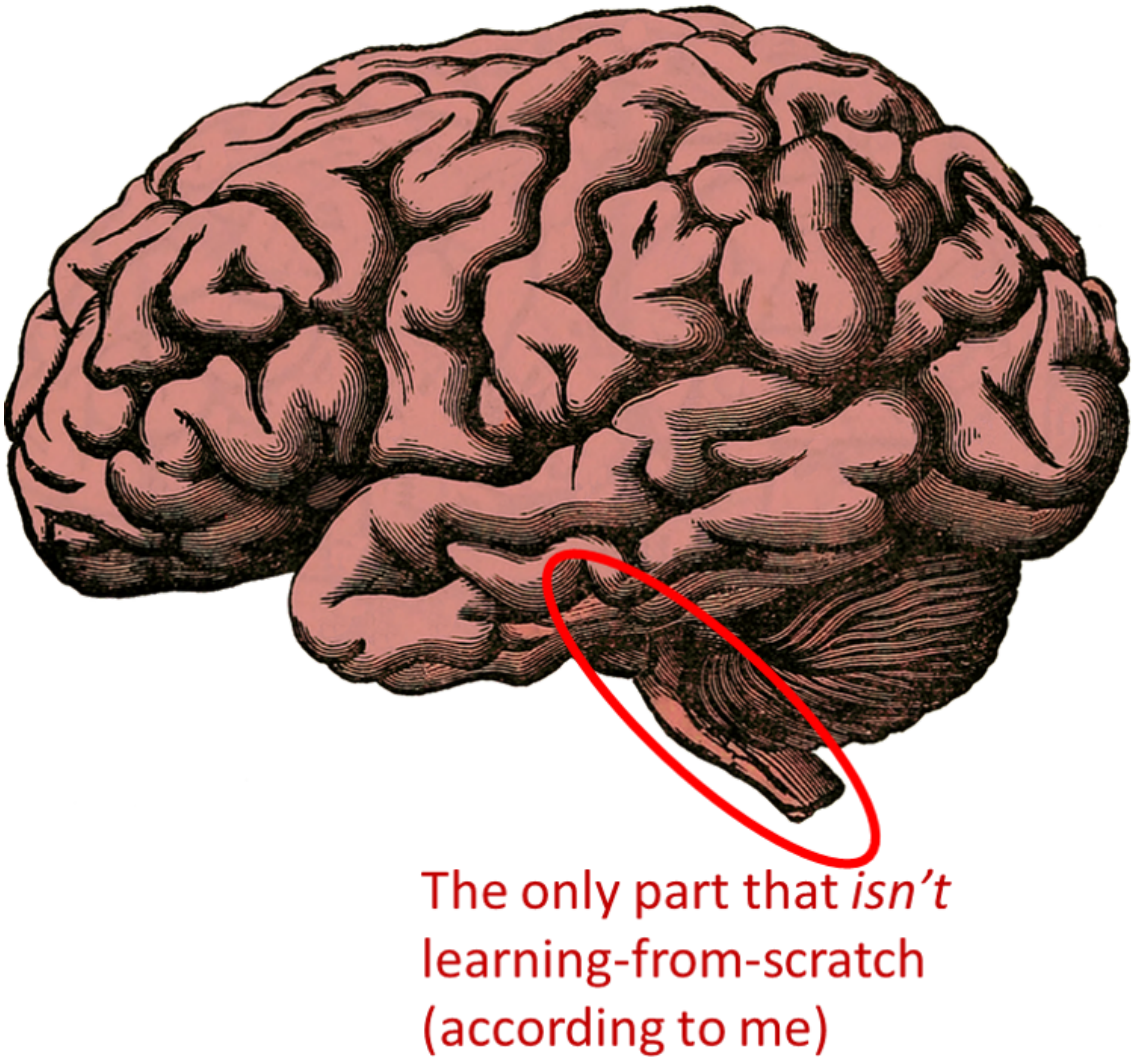
To be concrete, here are **three examples of “individual innate adjustable parameters” in the hypothalamus & brainstem:**

- I already mentioned the mouse hypothalamus circuit that says “if you keep winning fights, be more aggressive”—ref.
- Here’s a rat hypothalamus circuit that says “if you keep getting dangerously salt-deprived, increase your baseline appetite for salt”.
- The superior colliculus in the brainstem contains a visual map, auditory map, and saccade motor map, and it has a mechanism to keep all three lined up—so that when you see a flash or hear a noise, you immediately turn to look *in exactly the right*

direction. This mechanism involves plasticity—it can self-correct in animals wearing prism glasses, for example. I'm not familiar with the details, but I'm guessing it's something like: If you see a motion, and saccade to it, but the motion is not centered even *after* the saccade, then that generates an error signal that induces a corresponding incremental map shift. Maybe this whole system involves 8 adjustable parameters (scale and offset, horizontal and vertical, three maps to align), or maybe it's more complicated—again, I don't know the details.

See the difference? Go back to the table above if you're still confused.

2.4 My hypothesis: the telencephalon and cerebellum learn from scratch, the hypothalamus and brainstem don't



My hypothesis is that ~96% of the human brain by volume is running learning-from-scratch algorithms. The main exceptions

are the brainstem and hypothalamus, which together are around the size of your thumb. Image source

Three claims:

First, I think the **whole telencephalon** learns from scratch (and is useless at birth[1]). The telencephalon (a.k.a. "cerebrum") is mostly the neocortex in humans, plus the hippocampus, amygdala, most of the basal ganglia, and various more obscure bits and bobs.

Despite appearances, the model I like (due originally to the brilliant Larry Swanson) says that the whole telencephalon is organized into a nice three-layer structure (cortex, striatum, pallidum), and this structure aligns with a relatively small number of interconnected learning algorithms. See my (somewhat long and technical) post Big Picture of Phasic Dopamine for the details on that.

The **thalamus** is technically outside the telencephalon, but at least part of it is intimately interconnected with the cortex—some researchers describe it as functionally like an "extra layer" of cortex. So I would lump that part in with the learning-from-scratch telencephalon too.

The telencephalon and thalamus together comprise ~86% the volume of the human brain (ref).

Second, I think the **cerebellum also learns from scratch** (and is likewise useless at birth). The cerebellum is ~10% of adult brain volume (ref). More on the cerebellum in Post #4.

Third, I think the **hypothalamus and brainstem absolutely do NOT learn from scratch** (and they are very active and useful right from birth). I think other parts of the diencephalon are in the same category too—e.g. the habenula and pineal gland.

OK, that's my hypothesis.

I wouldn't be surprised if there were *minor* exceptions to this picture. Maybe there's some little nucleus somewhere in the telencephalon that orchestrates a biologically-adaptive behavior without first learning it from scratch. Sure, why not. But I currently think this picture is at least *broadly* right.

In the next two sections I'll talk about some evidence related to my hypothesis, and then what others in the field think of it.

2.5 Evidence on whether the telencephalon & cerebellum learn from scratch

2.5.1 Big-picture-type evidence

I find from reading and talking to people that the biggest sticking points against believing that the telencephalon and cerebellum learn from scratch is overwhelmingly *not* detailed discussion of neuroscience data etc. but rather:

1. failure to even consider this hypothesis as a possibility, and

2. confusion about the *consequences* of this hypothesis, and in particular how to flesh it out into a sensible big picture of brain and behavior.

If you've read this far, then #1 should no longer be a problem.

What about #2? A central type of question is *"If the telencephalon & cerebellum learn from scratch, then how do they do X?"*—for various different X. If there's an X for which we can't answer this question at all, it suggests that the learning-from-scratch hypothesis is wrong. Conversely, if we can find *really good* answers to this question for lots of X, it would offer evidence (though not proof) that the learning-from-scratch hypothesis is right. The upcoming posts in this series will, I hope, offer some of this type of evidence.

2.5.2 Neonatal evidence

If the telencephalon & cerebellum cannot produce biologically-adaptive outputs except by learning to do so over time, then it follows that any biologically-adaptive neonatal[1] behavior would have to be driven by the brainstem & hypothalamus. Is that right? It seems like the kind of thing that should be experimentally measurable, right? This 1991 paper indeed says "evidence has been accumulating that suggests that newborn perceptuomotor activity is mainly controlled by subcortical mechanisms". But I don't know if anything has changed in the 30 years since that paper—let me know if you've seen other references on this.

Actually, it's a harder question than it sounds. Suppose an infant does something biologically-adaptive…

- The first question we need to ask is: *really?* Maybe it's a bad (or wrongly-interpreted) experiment. For example, if an adult sticks his tongue out at a newborn infant human, will the infant stick out her own tongue as an act of imitation? Seems like a simple question, right? Nope, it's a decades-long raging controversy. A competing theory is centered around oral exploration: "tongue protrusion seems to be a general response to salient stimuli and is modulated by the child's interest in the stimuli"; a protruding adult tongue happens to elicit this response, but so do flashing lights and bursts of music. I'm sure some people know which newborn experiments are trustworthy, but I don't, at least not at the moment. And I'm feeling awfully paranoid after seeing two widely-respected books in the field (*Scientist in the Crib*, *Origin of Concepts*) repeat that claim about newborn tongue imitation as if it's a rock-solid fact.
- The second question we need to ask is: is it the result of within-lifetime learning? Remember, even a 3-month-old infant has had 4 million waking seconds of "training data" to learn from. In fact, even *zero*-day-old infants could have potentially been running learning-from-scratch algorithms in the womb.[1]
- The third question we need to ask is: what part of the brain is orchestrating this behavior? My hypothesis says that non-learned adaptive behaviors *cannot* be orchestrated by the telencephalon or cerebellum. But my hypothesis *does* allow such behaviors to be orchestrated by the brainstem! And figuring out which part of the neonatal brain is causally upstream of some behavior can be experimentally challenging.

2.5.3 "Uniformity" evidence

The "cortical uniformity" hypothesis says that every part of the neocortex runs a more-or-less similar algorithm. (…With various caveats, especially related to the non-uniform neural architecture and hyperparameters). Opinions differ on whether (or to what extent) cortical uniformity is true—I have a brief discussion of the evidence and arguments here (and links to more). I happen to think it's very probably true, at least in the weak sense that a future researcher who has a really good nuts-and-bolts understanding of how Neocortex Area #147 works would be *well on their way* to understanding how literally any other part of the

neocortex works. I won't be diving into that here; I consider it generally off-topic for this series.

I bring this up because *if* you believe in cortical uniformity, then you should *probably* believe in cortical learning-from-scratch as well. The argument goes as follows:

The adult neocortex does lots of apparently very different things: vision processing, sound processing, motor control, language, planning, and so on. How would one reconcile this fact with cortical uniformity?

Learning-from-scratch offers a plausible way to reconcile them. After all, we know that a single learning-from-scratch algorithm, fed with very different input data and supervisory signals, can wind up doing very different things—consider how transformer-architecture deep neural networks can be trained to generate natural-language text, or images, or music, or robot motor control signals, etc.

By contrast, if we *accept* cortical uniformity but *reject* learning-from-scratch, well, umm, I can't see any way to make sense of how that would work.

Analogously (but less often discussed than the neocortex case), should we believe in "allocortical uniformity"? As background, allocortex seems to be a simpler version of neocortex, with three layers instead of six; before the neocortex evolved, early amniotes are believed to have had 100% allocortex. Allocortex, like neocortex, does various different things: in adult humans, the hippocampus is involved in navigation and episodic memory, while the piriform cortex is involved in olfactory processing. So there's a potential analogous argument for learning-from-scratch there.

Moving on, I mentioned above (and more in Big Picture of Phasic Dopamine, and also upcoming in Post #5, Section 5.4.1) the idea (due to Larry Swanson) that the whole telencephalon seems to be organized into three layers—"cortex", "striatum", and "pallidum". I just talked about cortex; what about "striatal uniformity" and "pallidal uniformity"? Don't expect to find a dedicated literature review—in fact, the previous sentence seems to be the first time those two terms have ever been written down. But there are in fact at least some commonalities across each of those layers—e.g., medium spiny neurons exist everywhere in the striatum layer, I think. And I continue to believe that the picture I outlined in Big Picture of Phasic Dopamine (and upcoming Posts #5-#6) is a reasonable first-pass reconciliation between "everything we know about the striatum and pallidum" on the one hand, and "several variations on a certain learning-from-scratch algorithm" on the other hand.

In the cerebellum case, there is at least *some* literature on the uniformity hypothesis (search for the term "universal cerebellar transform"), but again no consensus. The adult cerebellum is likewise involved in apparently-different functions like motor coordination, language, cognition, and emotions. I personally believe in uniformity there too, with details coming up in Post #4.

2.5.4 Locally-random pattern separation

This is another reason that I personally put a lot of stock in the learning-from-scratch telencephalon and cerebellum. It's kinda specific, but very salient in my mind; see if you buy it.

2.5.4.1 What is pattern separation?

There is a common motif in the brain called "pattern separation". Let me explain what it is and why it exists.

Suppose you're an ML engineer working for a restaurant chain. Your boss tells you to predict sales for different candidate franchise locations.

The first thing you might do is to gather a bunch of data-streams—local unemployment rate, local restaurant ratings, local grocery store prices, whether there happens to be a novel coronavirus spreading around the world right now, etc. I like to call these "context data". You would use the context data as *inputs* to a neural network. The *output* of the network is supposed to be a prediction of the restaurant sales. You adjust the neural network weights (using supervised learning, with data from existing restaurants) to make that happen. No problem!

Pattern separation is when you add an extra step at the beginning. You take your various context data-streams, and *randomly* combine them in lots and lots of different ways. Then you sprinkle in some nonlinearity, and *voilà*! You now have *way more* context data-streams than you started with! Then *those* can be the inputs to the trainable neural net.[4]

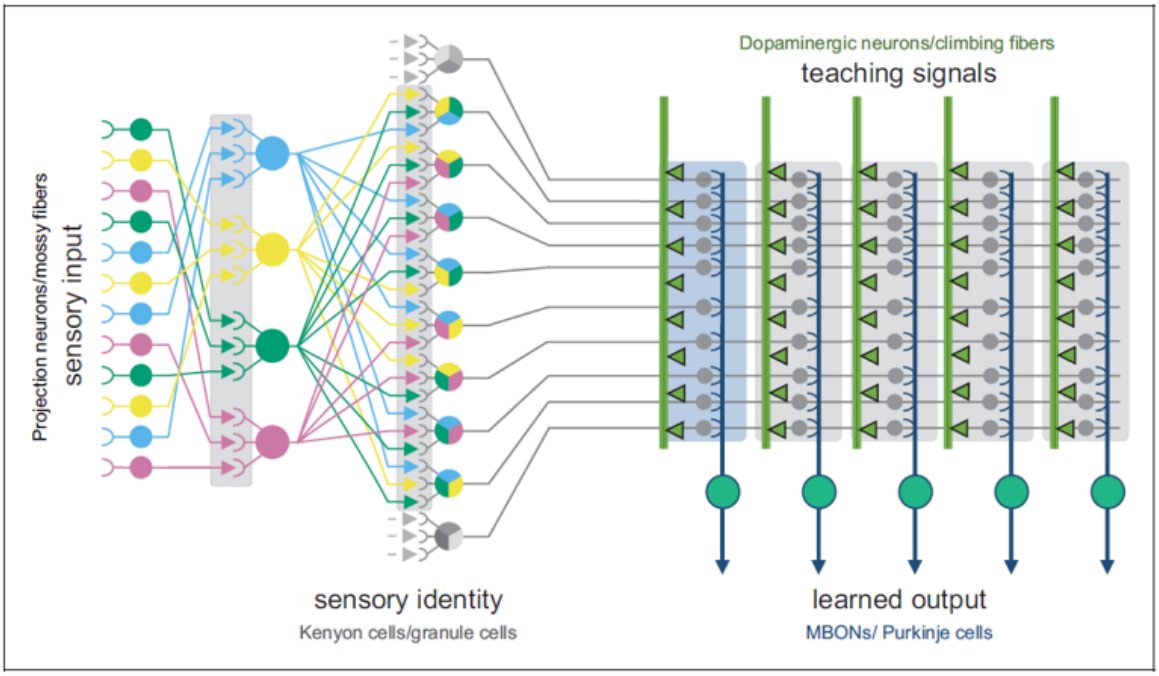


Illustration of (part of) fruit fly sensory processing. The tall vertical gray bar in the center-left is the "pattern separation" layer; it takes the organized sensory signals coming from the left, and remixes them up into a large number of different, (locally) random combinations. These are then sent rightward, to serve as "context" inputs for the supervised learning modules. Image source: Li et al. 2020.

In ML terms, pattern separation is like adding a very wide hidden layer, at the input side, with fixed weights. If the layer is wide enough, you'll find that some neurons in the layer are carrying useful representations, just by good luck. And then the next layer can use those useful neurons, and ignore the rest.

ML readers are now thinking to themselves: "OK, fine, but this is kinda dumb. Why add a extra-wide hidden layer at the beginning, with non-learnable weights? Why not just add a *normal*-sized hidden layer at the beginning, with *learnable* weights? Wouldn't that be easier and better?" Umm, probably! At least, it would indeed probably be better in this particular example.[5]

So why add a pattern-separation layer, instead of an extra learnable layer? Well, remember that in biological neurons, doing backprop (or something equivalent) through multiple learnable layers is *at best* a complicated procedure, and at worst totally impossible. Or at

least, that's my current impression. Backprop *as such* is widely acknowledged to be impossible to implement in biological neurons (cf. "the weight transport problem"). Various groups in ML and neuroscience have taken that as a challenge, and devised roughly 7 zillion different mechanisms that are (allegedly) biologically plausible and (allegedly) wind up functionally similar to backprop for one reason or another.[6] I haven't read all these papers. But anyway, even if it's *possible* to propagate errors through 2 learnable layers of biological neurons (or 3 layers, or even *N* layers), let's remember that it's an absolute *breeze* to do error-driven learning with just one learnable layer, using biological neurons. (All it takes is a set of identical synapses, getting updated by a 3-factor learning rule. Details coming up in Post #4.) So it's not crazy to think that evolution might settle on a solution like pattern separation, which gets *some* of the advantage of an extra learnable layer, but without the complication of actually propagating errors through an extra learnable layer.

2.5.4.2 *Where* is pattern-separation?

Pattern separation is thought to occur in a number of places, particularly involving the tiny and numerous neurons called "granule cells":

- The cerebellum has pattern-separating granule cells in its "granular layer" (ref). And boy are there a lot of them! Adult humans have 50 billion of them—more than half the neurons in your entire brain.
- The hippocampus has pattern-separating granule cells in its "dentate gyrus".
- The neocortex has pattern-separating granule cells in "layer 4", its primary (feedforward) input layer. To be clear, some neocortex is called "agranular", meaning that it lacks this granular layer. But that's just because not *all* the neocortex is processing inputs of the type that gets pattern-separated. Some neocortex is geared to outputs instead (details here).
- The fruit fly nervous system has a "mushroom body" consisting of "kenyon cells" which are also believed to be pattern-separators—see references here.

2.5.4.3 Why does pattern separation suggest learning-from-scratch?

The thing is, pattern separation seems to be a **locally random** process. What does "locally" mean here? Well, it's generally not true that *any* one input is equally likely to be mixed with *any* other input. (At least, not in fruit flies.) I only claim that it involves randomness at a small scale—like, out of *this* microscopic cluster of dozens of granule cells, *exactly* which cell connects with *exactly* which of the nearby input signals? I think the answer to those kinds of questions is: it's random.

Why do I say that it's probably (locally) random? Well, I can't be sure, but I do have a few reasons.

- From an algorithm perspective, (local) randomness seems like it would work, and indeed has some nice properties, like statistical guarantees about low overlap between sparse activation patterns.
- From an information-theory perspective, if there are 50 billion granule cells in an adult cerebellum, I find it pretty hard to imagine that the exact connections to each of them is deterministically orchestrated by the <1GB genome, while still satisfying the various algorithmic and biological constraints.
- From an experimental perspective, I'm not sure about vertebrates, but at least in the fruit fly case, genetically-identical fruit flies are known to have different kenyon cell connectivity (ref).

Anyway, if pattern-separation is a (locally) random process, then that means that *you can't do anything useful with the outputs of a pattern-separation layer, except by learning to do*

so. In other words, we wind up with a learning-from-scratch algorithm! (Indeed, one that would *stay* learning-from-scratch even in the face of evolutionary pressure to micromanage the initial parameters!)

2.5.5 Summary: I don't pretend that I've *proven* the hypothesis of learning-from-scratch telencephalon and cerebellum, but I'll ask you to suspend disbelief and read on

In my own head, when I mash together all the considerations I discussed above (big-picture stuff, neonatal stuff, uniformity, and locally-random pattern separation), I wind up feeling quite confident in my hypothesis of a learning-from-scratch telencephalon and cerebellum. But really, everything here is suggestive, not the kind of definitive, authoritative discussion that would convince every skeptic. The comprehensive scholarly literature review on learning-from-scratch telencephalon and cerebellum, so far as I know, has yet to be written.

Don't get me wrong; I would *love* to write that! I would dive into all the relevant evidence, like everything discussed above, plus other things like experiments on decorticate rats, etc. That would be awesome, and I may well do that at some point in the future. (Or reach out if you want to collaborate!)

But meanwhile, I'm going to treat the hypothesis as if it were true. This is just for readability—the whole rest of the series will be exploring bigger-picture consequences of the hypothesis, and it would get really annoying if I put apologies and caveats in every other sentence.

2.6 Is my hypothesis consensus, or controversial?

Weirdly, I just don't know! This is *not* a hot topic of discussion in neuroscience. I think most people haven't even thought to formulate "what parts of the brain learn from scratch" as an explicit question, let alone a question of absolutely central importance.

(I heard from an old-timer in the field that the question "what parts of the brain learn from scratch?" smells too much like "nature vs nurture". According to them, everyone had a fun debate about "nature vs nurture" in the 1990s, and then they got sick of it and moved on to other things! Indeed, I asked for a state-of-the art reference on the evidence for learning-from-scratch in the telencephalon and cerebellum, and they suggested a book from 25 years ago! It's a good book—in fact I had already read it. But *c'mon*!! We've learned new things since 1996, right??)

Some data points:

- Neuroscientist Randall O'Reilly explicitly endorses a learning-from-scratch neocortex (in agreement with me). He talks about it here (30:00), citing this paper on infant face recognition as a line of evidence. In fact, I think O'Reilly would agree with at least *most* of my hypothesis, and maybe all of it.
- I'm also pretty confident that Jeff Hawkins and Dileep George would endorse my hypothesis, or at least something very close to it. More on them in the next post.
- A commenter suggested the book *Beyond Evolutionary Psychology* by George Ellis & Mark Solms (2018), which (among other things) argues for something strikingly similar to my hypothesis—they list the brain's "soft-wired domains" as consisting of the neocortex, the cerebellum, and "parts of the limbic system, for instance most of the

hippocampus and amygdala, and large parts of the basal ganglia" (page 209). Almost a perfect match to *my* list! But their notion of "soft-wired domains" is defined somewhat differently than my notion of "learning from scratch", and indeed I disagree with the book in numerous areas. But anyway, the book has lots of relevant evidence and literature. Incidentally, the book was mainly written as an argument against an "innate cognitive modules" perspective exemplified by Steven Pinker's *How the Mind Works* (1994). So it's no surprise that Steven Pinker would disagree with the claim that the neocortex learns from scratch (well, I'm 99% confident that he would)—see, for example, his book *The Blank Slate* (2003) chapter 5.
- Some corners of computational neuroscience—particularly those with ties to the deep learning community—seem very enthusiastic about learning-from-scratch algorithms in the brain *in general*. But the discourse there doesn't seem specific enough to answer my question. For example, I'm looking for statements like "Within-lifetime learning algorithms are a good starting point for understanding the neocortex, but a bad starting point for understanding the medulla." I can't find anything like that. Instead I see, for example, the paper "A deep learning framework for neuroscience" (by 32 people including Blake Richards and Konrad Kording), which says something like "Learning algorithms are very important for the brain, and sometimes those learning algorithms are within-lifetime learning algorithms, whereas other times the only learning algorithm is evolution." But which parts of the brain are in which category? The paper doesn't say.
- My vague impression from sporadically reading papers with computational models of the neocortex, hippocampus, cerebellum, and striatum, from various different groups, is that the models are at least *often* learning-from-scratch models, but not always.

In summary, while I'm uncertain, there's some reason to believe that my hypothesis is not too far outside the mainstream…

But it's only "not too far outside the mainstream" in a kind of tunnel-vision sense. Almost nobody in neuroscience is taking the hypothesis seriously enough to grapple with its *bigger-picture consequences*. As mentioned above, if you believe (as I do) that "if the telencephalon or cerebellum perform a useful function X, they must have *learned* to perform that function, within the organism's lifetime, somehow or other", then that immediately spawns a million follow-up questions of the form: "*How* did it learn to do X? Is there a ground-truth that it's learning from? What is it? Where does it come from?" I have a hard time finding good discussions of these questions in the literature. Whereas I'm asking these questions *constantly*, as you'll see if you read on.

In this series of posts, I'm going to talk extensively about the bigger-picture framework around learning-from-scratch. By contrast, I'm going to talk relatively *little* about the nuts-and-bolts of how the learning algorithms work. That would be a complicated story which is not particularly relevant for AGI safety. And at least in some cases, nobody really knows the exact learning algorithms anyway.

2.7 Why does learning-from-scratch matter for AGI safety?

Much more on this later, but here's a preview.

The single most important question in AGI safety is: Is the AGI trying to do something that we didn't intend for it to be trying to do?

If no, awesome! This is sometimes called "intent alignment". Granted, even with intent alignment, we can't quite declare *complete* victory over accident risk—the AGI can still screw things up despite good intentions (see Post #11). But we've made a lot of progress, and probably averted the worst problems.

By contrast, if the AGI *is* trying to do something that we hadn't intended for it to be trying to do, that's where we get into the *really bad* minefield of catastrophic accidents. And as we build more and more capable AGIs over time, the accidents get worse, not better, because the AGI will become more skillful at figuring out *how best* to do those things that we had never wanted it to be doing in the first place.

So the critical question is: how does the AGI wind up trying to do one thing and not another? And the follow-up question is: if we want the AGI to be trying to do a particular thing X (where X is "act ethically", or "be helpful", or whatever—more on this in future posts), what code do we write?

Learning-from-scratch means that the AGI's common-sense world-model involves one or more big data structures that are built from scratch during the AGI's "lifetime" / "training". The stuff inside those data structures is not necessarily human-interpretable[7]—after all, it was never put in by humans in the first place!

And unfortunately, the things that we want the AGI to be trying to do—"act ethically", or "solve Alzheimer's", or whatever—are naturally defined in terms of abstract concepts. At best, those concepts are buried somewhere inside those big data structures. At worst (e.g. early in training), the AGI might not even have those concepts in the first place. So how do we write code such that the AGI wants to solve Alzheimer's?

In fact, evolution has the same problem! Evolution would *love* to paint the abstract concept "Have lots of biological descendants" with positive valence, but thanks to learning-from-scratch, the genome doesn't know which precise set of neurons will ultimately be representing this concept. (And not all humans have a "biological descendants" concept anyway.) The genome does other things instead, and later in the series I'll be talking more about what those things are.

2.8 Timelines-to-brain-like-AGI part 1/3: how hard will it be to reverse-engineer the learning-from-scratch parts of the brain, well enough for AGI?

This isn't *exactly* on-topic, so I don't want to get too deep into it. But in Section 1.5 of the previous post, I mentioned that there's a popular idea that "brain-like AGI" (as defined in Section 1.3.2 of the previous post) is probably centuries away because the brain is so very horrifically complicated. I then said that I strongly disagreed. Now I can say a bit about why.

As context, we can divide the "build brain-like AGI" problem up into three pieces:

1. Reverse-engineer the learning-from-scratch parts of the brain (telencephalon & cerebellum) well enough for AGI,
2. Reverse-engineer everything else (mainly the brainstem & hypothalamus) well enough for AGI,
3. Actually build the AGI—including hardware-accelerating the code, running model trainings, working out all the kinks, etc.

This section is about #1. I'll get back to #2 & #3 in Sections 3.7 & 3.8 of the next post.

Learning-from-scratch is highly relevant here because **reverse-engineering learning-from-scratch algorithms is a *way* simpler task than reverse-engineering trained models**.

For example, think of the OpenAI Microscope visualizations of different neurons in a deep neural net. There's so much complexity! But no human needed to design that complexity; it was automatically discovered by the learning algorithm. The learning algorithm itself is comparatively simple—gradient descent and so on.

Here are some more intuitions on this topic:

- I think that learning-from-scratch algorithms kinda *have* to be simple, because they have to draw on broadly-applicable regularities—"patterns tend to recur", and "patterns are often localized in time and space", and "things are often composed of other things", and so on.
- Human brains were able to invent quantum mechanics. I can *kinda* see how a learning algorithm based on simple, general principles like "things are often composed of other things" (as above) can eventually invent quantum mechanics. I *can't* see how a horrifically-complicated-Rube-Goldberg-machine of an algorithm can invent quantum mechanics. It's just so wildly different from anything in the ancestral environment.
- The learning algorithm's neural architecture, hyperparameters, etc. could be kinda complicated. I freely admit it. For example, this study says that the neocortex has 180 architecturally-distinguishable areas. But on the other hand, future researchers don't need to reinvent that stuff from scratch; they could also just "crib the answers" from the neuroscience literature. And also, not *all* that complexity is *necessary* for human intelligence—as we know from the ability of infants to (sometimes) fully recover from various forms of brain damage. Some complexity might just help speed the learning process up a bit, on the margin, or might help with unnecessary-for-AGI things like our sense of smell.
- In Section 2.5.3, I discussed the "cortical uniformity" hypothesis and its various cousins. If true, it would greatly limit the potential difficulty of understanding how those

parts of the brain work. But I don't think anything I'm saying here depends on the "uniformity" hypotheses being true, let alone *strictly* true.

Going back to the question at issue. **In another (say) 20 years, will we understand the telencephalon and cerebellum well enough to build the learning-from-scratch part of an AGI?**

I say: I don't know! Maybe we will, maybe we won't.

There are people who disagree with me on that. *They* claim that the answer is "Absolutely 100% not! Laughable! How dare you even *think* that? That's the kind of thing that only a self-promoting charlatan would say, as they try to dupe money out of investors! That's not the kind of thing that a serious cautious neuroscientist would say!!!" Etc. etc.

My response is: I think that this is wildly-unwarranted overconfidence. I don't see any good reason to rule out figuring this out in (say) 20 years, or even 5 years. Or maybe it *will* take 100 years! I think we should remain uncertain. As they say, "Predictions are hard, especially about the future."

1. ^

   I keep saying that "learning from scratch" implies "unhelpful for behavior *at birth*". This is an oversimplification, because it's possible for "within-lifetime learning" to happen in the womb. After all, there should already be *plenty* of data to learn from in the womb—interoception, sounds, motor control, etc. And maybe retinal waves too—those could be functioning as fake sensory data for the learning algorithm to learn from.

2. ^

   Minor technicality: Why did I say the input data and supervisory signals for the neocortex (for example) come from *outside* the neocortex? Can't one part of the neocortex get input data and/or supervisory signals from a different part of the neocortex? Yes, of course. However, I would instead describe that as "part of the neocortex's neural architecture". By analogy, in ML, people normally would NOT say "ConvNet-layer-12 gets input data from ConvNet-layer-11". Instead, they would be more likely to say "The ConvNet (as a whole) gets input data from outside the ConvNet". This is just a way of talking, it doesn't really matter.

3. ^

   I'm framing this as a "made-up example" because I'm trying to make a simple conceptual point, and don't want to get bogged down in complicated uncertain empirical details. That said, the bird song thing is not *entirely* made up—it's at least "inspired by a true story". See discussion here of Gadagkar 2016, which found that a subset of dopamine neurons in the songbird brainstem send signals that look like RL rewards for song quality, and those signals go specifically to the vocal motor system, presumably training it to sing better. The missing part of that story is: what calculations are upstream of those particular dopamine neurons? In other words, how does the bird brain judge its own success at singing? For example, does it match its self-generated auditory inputs to an innate template? Or maybe the template is generated in a more complicated way—say, involving listening to adult birds of the same species? Or something else? I'm not sure the details here are known—or at least, I don't personally know them.

4. ^

   Why is it called "pattern separation"? It's kinda related to the fact that a pattern-separator has more output lines than input lines. For example, you might regularly

encounter five different “patterns” of sensory data, and maybe all of them consist of activity in the same set of 30 input lines, albeit with subtle differences—maybe one pattern has such-and-such input signal slightly stronger than in the other patterns, etc. So on the input side, we might say that these five patterns “overlap”. But on the *output* side, maybe these five patterns would wind up activating entirely different sets of neurons. Hence, the patterns have been “separated”.

5. ^

   In other examples, I think pattern separation is serving other purposes too, e.g. sparsifying the neuron activations, which turns out to be very important for various reasons, including not getting seizures.

6. ^

   If you want to dive into the rapidly-growing literature on biologically-plausible backprop-ish algorithms, a possible starting point would be References #12, 14, 34–38, 91, 93, and 94 of A deep learning framework for neuroscience.

7. ^

   There is a field of “machine learning interpretability”, dedicated to interpreting the innards of learned-from-scratch “trained models”—example. I (along with everyone else working on AGI safety) strongly endorse efforts to advance that field, including tackling much bigger models, and models trained by a wider variety of different learning algorithms. Also on this topic: I sometimes hear an argument that a brain-like AGI using a brain-like learning algorithm will produce a relatively more human-interpretable trained model than alternatives. This strikes me as maybe true, but far from guaranteed, and anyway “relatively more human-interpretable” is different than “very human-interpretable”. Recall that the neocortex has ~100 trillion synapses, and an AGI could eventually have many more than that.

[Intro to brain-like-AGI safety] 3. Two subsystems: Learning & Steering

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

3.1 Post summary / Table of contents

Part of the "Intro to brain-like-AGI safety" post series .

In the previous post I defined the notion of "learning from scratch" algorithms—a broad category that includes, among other things, any randomly-initialized machine learning algorithm (no matter how complicated), and any memory system that starts out empty. I then proposed a division of the brain into two parts based on whether or not they learn from scratch. Now I'm giving them names:

The **Learning Subsystem** is the 96% of the brain that "learns from scratch"—basically the telencephalon and cerebellum.

The **Steering Subsystem** is the 4% of the brain that *doesn't* "learn from scratch"—basically the hypothalamus and brainstem.

(See previous post for a more detailed anatomical breakdown.)

This post will be a discussion of this two-subsystems picture in general, and of the Steering Subsystem in particular.

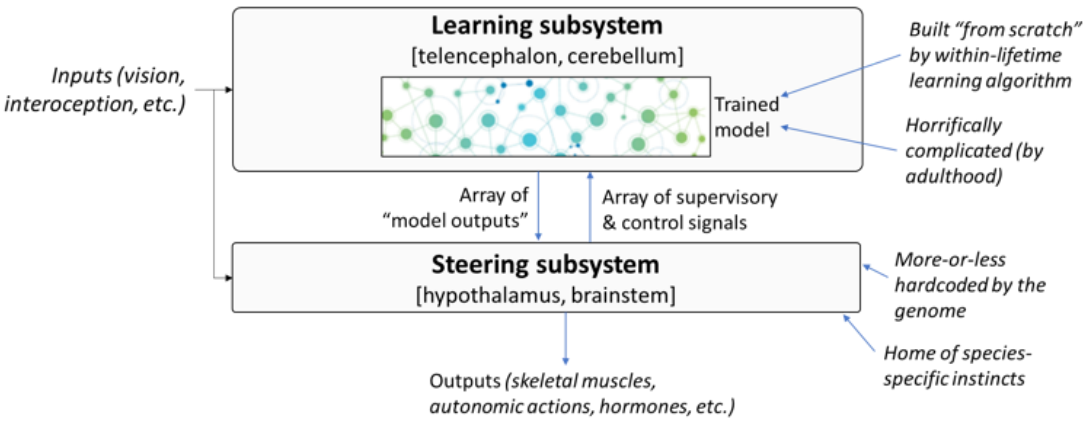
- In Section 3.2, I'll talk about the big picture of what these subsystems do and how they interact. As an example, I'll explain why each subsystem needs its own sensory-processing circuitry—for example, why visual inputs get processed by *both* the visual cortex in the Learning Subsystem, *and* the superior colliculus in the Steering Subsystem.
- In Section 3.3, I'll acknowledge that this two-subsystem picture has some echoes of the discredited "triune brain theory". But I'll argue that the various problems with triune brain theory do not apply to my two-subsystem picture.
- In Section 3.4, I'll discuss three categories of ingredients that could go into a Steering Subsystem:
  - Category A: Things that are plausibly essential for general intelligence (e.g. an innate drive for curiosity),
  - Category B: Everything else in the human steering subsystem (e.g. an innate drive to be kind to your friends),
  - Category C: Any other possibility that an AGI programmer might dream up, even if it's radically different from anything in humans or animals (e.g. an innate drive to correctly predict stock prices).
- In Section 3.5, I'll relate those categories to how I expect people to build brain-like AGIs, arguing that *"brain-like AGIs with radically non-human (and dangerous) motivations"* is not an oxymoron; rather, it's the default expected outcome, unless we work to prevent it.
- In Section 3.6, I'll discuss the fact that Jeff Hawkins has a two-subsystems perspective similar to mine, yet argues *against* AGI catastrophic accidents being a risk. I'll say where I think he goes wrong.
- Sections 3.7 and 3.8 will be the final two parts of my "timelines to brain-like AGI" discussion. The first part was Section 2.8 in the previous post, where I argued that reverse-engineering the Learning Subsystem (at least well enough to enable brain-like AGI) is something that could plausibly happen soon, like within the next decade or two, although it could also take longer. Here, I'll complete that story by arguing that this same thing is true of reverse-engineering the Steering Subsystem (at least well enough to enable brain-like AGI), and of getting the algorithms cleaned up and scaled up, running model trainings, and so on.

- Section 3.9 is a quick non-technical discussion on the wildly divergent attitudes that different people take towards the timeline to AGI, even when they agree on the probabilities. For example, you can have two people agree that the odds are 3:1 against having AGI by 2042, but one might emphasize how low that probability is ("You see? AGI probably isn't going to arrive for *decades*"), while the other might emphasize how *high* that probability is. I'll talk a bit about the factors that can underlie those attitudes.

3.2 Big picture

In the last post, I claimed that 96% of the brain by volume—roughly the telencephalon (neocortex, hippocampus, amygdala, most of the basal ganglia, and a few other things) and cerebellum—"learns from scratch", in the sense that early in life its outputs are all random garbage, but over time they become extremely helpful thanks to within-lifetime learning. (More details and caveats in the previous post.) I'm now calling this part of the brain the **Learning Subsystem**.

The rest of the brain—mainly the brainstem and hypothalamus—I'm calling the **Steering Subsystem**.



How are we supposed to think about these?

Let's start with the Learning Subsystem. As discussed in the last post, this subsystem has some interconnected, innate learning algorithms, with innate neural architectures and innate hyperparameters. It also has *lots* (as in billions or trillions) of adjustable parameters of some sort (usually assumed to be synapse strength, but this is controversial and I won't get into it), and the values of these parameters start out random. The Learning Subsystem's algorithms thus emit random unhelpful-for-the-organism outputs at first—for example, perhaps they cause the organism to twitch. But over time, various supervisory signals and corresponding update rules sculpt the values of the system's adjustable parameters, tailoring them within the animal's lifetime to do tricky biologically-adaptive things.

Next up: the Steering Subsystem. How do we think intuitively about that one?

First off, imagine a repository with lots of species-specific instincts and behaviors, all hardcoded in the genome:

- *"In order to vomit, contract muscles A,B,C, and release hormones D,E,F."*
- *"If sensory inputs satisfy the thus-and-such heuristics, then I am probably eating something healthy and energy-dense; this is good and I should react by issuing signals G,H,I."*
- *"If sensory inputs satisfy the thus-and-such heuristics, then I am probably leaning over a precipice; this is bad and I should react by issuing signals J,K,L."*
- *"When I'm cold, get goosebumps."*

- *"When I'm under-nourished, do the following tasks: (1) emit a hunger sensation, (2) start rewarding the neocortex for getting food, (3) reduce fertility and growth, (4) reduce pain sensitivity, etc." (ref).*

An especially-important task of the Steering Subsystem is sending supervisory and control signals to the Learning Subsystem. Hence the name: the Steering Subsystem *steers* the learning algorithms to do adaptive things.

For example: How is it that a *human* neocortex learns to do adaptive-for-a-*human* things, while a *squirrel* neocortex learns to do adaptive-for-a-*squirrel* things, if they're both vaguely-similar learning-from-scratch algorithms?

The main part of the answer, I claim, is that the learning algorithms get "steered" differently in the two cases. An especially important aspect here is the "reward" signal for reinforcement learning. You can imagine that the human brainstem sends up a "reward" for achieving high social status, whereas the squirrel brainstem sends up a "reward" for burying nuts in the fall. (This is oversimplified; I'll be elaborating on this story as we go.)

By the same token, in ML, the *same* learning algorithm can get really good at playing chess (given a certain reward signal and sensory data) *or* can get really good at playing Go (given a *different* reward signal and sensory data).

To be clear, despite the name, "steering" the Learning Subsystem is but one task of the Steering Subsystem. The Steering Subsystem can also just up and do things, all by itself, without any involvement from the Learning Subsystem! This is a good plan if doing those things is important right from birth, or if messing them up even once is fatal. An example I mentioned in the last post is that mice apparently have a brainstem bird-detecting circuit wired directly to a brainstem running-away circuit.

An important dynamic to keep in mind is that the brain's Steering Subsystem can*not* directly access our common-sense understanding of the world. For example, the Steering Subsystem can implement reactions like "when eating, manufacture digestive enzymes". But as soon as we start talking about the abstract concepts that we use to navigate the world—grades, debt, popularity, soy sauce, and so on—we have to assume that the Steering Subsystem has no idea what any of things are, unless we can come up with some story for how it found out. And sometimes there *is* such a story! We'll see a lot of those kinds of stories as we go, particularly Post #7 (for a simple example of wanting to eat cake) and Post #13 (for the trickier case of social instincts).

3.2.1 Each subsystem generally needs its own sensory processor

For example, in the case of vision, the Steering Subsystem has its superior colliculus, while the Learning Subsystem has its visual cortex. For taste, the Steering Subsystem has its gustatory nucleus of the medulla, while the Learning Subsystem has its gustatory cortex. Etc.

Isn't that redundant? Some people think so! The book *Accidental Mind* by David Linden cites the existence of two sensory-processing systems as a beautiful example of kludgy brain design resulting from evolution's lack of foresight. But I disagree. They're not redundant. If I were making an AGI, I would *absolutely* put in two sensory-processing systems!

Why? Suppose that Evolution wants to build a reaction circuit where a genetically-hardwired sensory cue triggers a genetically-hardwired response. For example, as mentioned above, if you're a mouse, then an expanding dark blob in the upper field-of-view often indicates an incoming bird, and therefore the mouse genome hardwires an expanding-dark-blob-detector to a running-away behavioral circuit.

And I claim that, when building this reaction, the genome *cannot use the visual cortex as its expanding-dark-blob-detector*. Why not? Remember the previous post: the visual cortex learns

from scratch! It takes unstructured visual data and builds a predictive model around it. You can (loosely) think of the visual cortex as a scrupulous cataloguer of patterns in the inputs, and of patterns in the patterns in the inputs, etc. One of these patterns might correspond to expanding dark blobs in the upper field-of-view. Or maybe not! And even if one does, the genome doesn't know in advance *which precise neurons* will be storing that particular pattern. And thus, the genome cannot hardwire those neurons to the running-away behavioral controller.

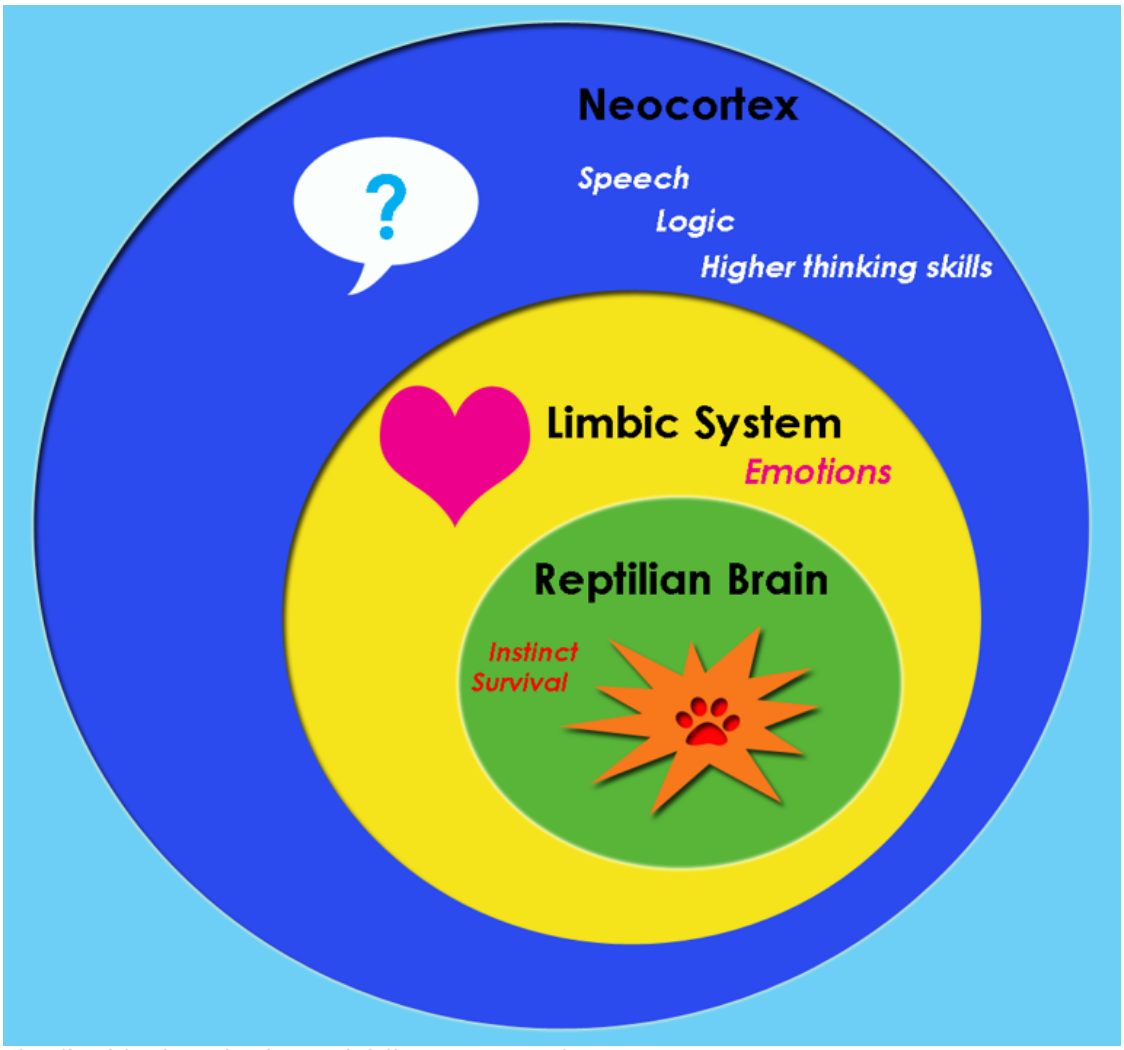
So in summary:

- *Building sensory processing into the Steering Subsystem is a good idea*, because there are lots of areas where it's highly adaptive to attach a genetically-hardwired sensory cue to a corresponding reaction. In the human case, think of fear-of-heights, fear-of-snakes, aesthetics-of-potential-habitats, aesthetics-of-potential-mates, taste-of-nutritious-food, sound-of-screaming, feel-of-pain, and on and on.
- *Building sensory processing into the Learning Subsystem is also a good idea*, because using learning-from-scratch algorithms to learn arbitrary predictive patterns in sensory input within a lifetime is, well, *a really good idea*. After all, many useful sensory patterns are hyper-specific—e.g. "the smell of this one specific individual tree"—such that a corresponding hardwired sensory pattern detector could not have evolved.

Thus, the brain's two sensory-processing systems is *not* an example of kludgy design. It's an example of Orgel's Second Rule: "evolution is cleverer than you are"!

3.3 "Triune Brain Theory" is wrong, but let's not throw out the baby with the bathwater

In the 1960s & 70s, Paul MacLean & Carl Sagan invented and popularized an idea called the Triune Brain. According to this theory, the brain consists of three layers, stacked on top of each other like an ice cream cone, and which evolved in sequence: first the "lizard brain" (a.k.a. "old brain" or "reptilian brain") closest to the spinal cord (consisting of the brainstem and basal ganglia); second the "limbic system" wrapped around that (consisting of the amygdala, hippocampus, and hypothalamus), and finally, layered on the outside, the neocortex (a.k.a. "new brain")—the pièce de résistance, the pinnacle of evolution, the home of human intelligence!!!



The (bad!) triune brain model (image source).

Well, it's by now well known that **Triune Brain Theory is rubbish**. It lumps brain parts in a way that makes neither functional nor embryological sense, and the evolutionary story is profoundly wrong. For example, half a billion years ago, the earliest vertebrates already had the precursors of *all three* layers of the triune brain—including a "pallium" which would eventually (in our lineage) segregate into the neocortex, hippocampus, part of the amygdala, etc. (ref).

So yeah, Triune Brain Theory is rubbish. But I freely admit: the story I like (previous section) kinda *rings* of triune brain theory. My Steering Subsystem looks suspiciously like MacLean's "reptilian brain". My Learning Subsystem looks suspiciously like MacLean's "limbic system and neocortex". MacLean & I have some disagreements about *exactly* what goes where, and whether the ice cream cone has two scoops versus three. But there's definitely a resemblance.

My two-subsystem story in this post is not original. You'll hear a similar story from Jeff Hawkins, Dileep George, Elon Musk, and others.

But those other people tell this story *in the tradition of triune brain theory*, and in particular keeping its problematic aspects, like the "old brain" and "new brain" terminology.

There's no need to do that!! We can keep the two-subsystem story, while throwing out the triune brain baggage.

So my story is: I think that half a billion years ago, the earliest vertebrates had a (simpler!) learning-from-scratch algorithm in their (proto) telencephalon, and it was "steered" by supervisory signals from their (simpler, proto) brainstem and hypothalamus.

Indeed, we can go back even earlier than vertebrates! There seems to be a homology between the learning-from-scratch cortex in humans and the learning-from-scratch "mushroom body" in fruit flies! (Further discussion here.) I note, for example, that in fruit flies, odor signals go to both the mushroom body and the lateral horn, in beautiful agreement with the general principle that sensory inputs need to go to both the Learning Subsystem and the Steering Subsystem (Section 3.2.1 above).

Anyway, in the 700 million years since our last common ancestor with insects, *both* the Learning Subsystem *and* the Steering Subsystem have dramatically expanded and elaborated in our lineage.

But that doesn't mean that they contribute equally to "human intelligence". Again, both are essential, but I think it's strongly suggestive that ~96% of human brain volume is the Learning Subsystem. Focusing more specifically on the telencephalon part (which includes the neocortex in mammals), its fraction of brain volume is 87% in humans (ref), 79% in chimps (ref), 77% in certain parrots, 51% in chickens, 45% in crocodiles, and just 22% in frogs (ref). There's an obvious pattern here, and I think it's right: namely, that to get recognizably intelligent and flexible behavior, you need a massively-scaled-up Learning Subsystem.

See? I can tell my two-subsystem story with none of that "old brain, new brain" nonsense.

3.4 Three types of ingredients in a Steering Subsystem

I'll start with the summary table, and then elaborate on it in the following subsections.

3.4.1 Summary table

Category of Steering Subsystem ingredient	**Possible examples**	**Present in (competent) humans?**	**Expected in future AGIs?**
(A) Things the Steering Subsystem *needs* to do in order to get general intelligence	▪ Curiosity drive (?) ▪ Drive to attend to certain types of things in the environment (humans, language, technology, etc.) (?) ▪ General involvement in helping establish the Learning Subsystem neural architecture (?)	Yes, by definition	Yes
(B) Everything else in		Often, but not	

a neurotypical human's Steering Subsystem	· **Social instincts** (which underlie altruism, love, remorse, guilt, sense-of-justice, loyalty, etc.) · Drives underlying disgust, aesthetics, transcendence, serenity, awe, hunger, pain, fear-of-spiders, etc.	always—for example, high-functioning sociopaths seem to be missing some of the usual social instincts.	Not "by default", but it's *possible* if we: (1) figure out exactly how they work, and (2) convince AGI developers to put them in.
(C) Every other possibility, most of which are *completely unlike anything* in the Steering Subsystem of humans or indeed any animal	· Drive to increase a company's bank account balance? · Drive to invent a better solar cell? · Drive to do whatever my human supervisor wants me to do? *(There's a catch: no one knows how to implement this one!)*	No	Yes "by default". If something is a bad idea, we can try to convince AGI developers not to do that.

3.4.2 Aside: what do I mean by "drives"?

I'll elaborate on this picture in later posts, but for now let's just say that the Learning Subsystem does reinforcement learning (among other things), and the Steering Subsystem sends it rewards. The components of the reward function relate to what I'll call "innate drives"—they're the root cause of why some things are inherently motivating / appetitive and other things are inherently demotivating / aversive.

Explicit goals like "I want to get out of debt" are different from innate drives. Explicit goals come out of a complicated dance between "innate drives in the Steering Subsystem" and "learned content in the Learning Subsystem". Again, much more on that topic in future posts.

Remember, innate drives are in the Steering Subsystem, whereas the abstract concepts that make up your conscious world are in the Learning Subsystem. For example, if I say something like "altruism-related innate drives", you need to understand that I'm *not* talking about "the abstract concept of altruism, as defined in an English-language dictionary", but rather "some innate Steering Subsystem circuitry which is *upstream* of the fact that neurotypical people sometimes find altruistic actions to be inherently motivating". There is *some* relationship between the abstract concepts and the innate circuitry, but it might be a complicated one—nobody expects a one-to-one relation between *N* discrete innate circuits and a corresponding set of *N* English-language words describing emotions and drives.[1]

With that out of the way, let's move on to more details about that table above.

3.4.3 Category A: Things the Steering Subsystem *needs* to do in order to get general intelligence (e.g. curiosity drive)

Let's start with the **"curiosity drive"**. If you're not familiar with the background of "curiosity" in ML, I recommend *The Alignment Problem* by Brian Christian, chapter 6, which contains the gripping story of how researchers eventually got RL agents to win the Atari game *Montezuma's Revenge*. Curiosity drives seem essential to good performance in ML, and humans also seem to have an innate curiosity drive. I assume that future AGI algorithms will need a curiosity drive as well, or else they just won't work.

To be more specific, I think this is a bootstrapping issue—I think we need a curiosity drive early in training, but can probably turn it off eventually. Specifically, let's say there's an AGI that's generally knowledgeable about the world and itself, and capable of getting things done, and right now it's trying to invent a better solar cell. I claim it probably doesn't need to feel an innate curiosity drive. Instead it may seek new information, and seek surprises, *as if* it were innately curious, because it has learned through experience that seeking those things tends to be an effective strategy for inventing a better solar cell. In other words, something like curiosity can be *motivating as a means to an end*, even if it's not *motivating as an end in itself*—curiosity can be a learned metacognitive heuristic. See instrumental convergence. But that argument does not apply early in training, when the AGI starts from scratch, knowing nothing about the world or itself. Instead, early in training, I think we really need the Steering Subsystem to be holding the Learning Subsystem's hand, and pointing it in the right directions, if we want AGI.

Another possible item in Category A is an **innate drive to pay attention to certain things in the environment, e.g. human activities, or human language, or technology**. I don't know *for sure* that this is necessary, but it seems to me that a curiosity drive *by itself* wouldn't do what we want it to do. It would be completely undirected. Maybe it would spend eternity running Rule 110 in its head, finding deeper and deeper patterns, while completely ignoring the physical universe. Or maybe it would find deeper and deeper patterns in the shapes of clouds, while completely ignoring everything about humans and technology. In the human brain case, the human brainstem definitely has a mechanism for forcing attention onto human faces (ref), and I strongly suspect that there's a system that forces attention onto human speech sounds as well. I could be wrong, but my hunch is that something like that will need to be in AGIs too. As above, if this drive is necessary at all, it might only be necessary early in training.

What else might be in Category A? On the table above, I wrote the vague "General involvement in helping establish the Learning Subsystem neural architecture". This includes sending reward signals and error signals and hyperparameters etc. to particular parts of the neural architecture in the Learning Subsystem. For example, in Post #6 I'll talk about how only *part* of the neural architecture gets the main RL reward signal. I think of these things as (one aspect of) how the Learning Subsystem's neural architecture is actually implemented. AGIs will have some kind of neural architecture too, although maybe not exactly the same as humans'. Therefore, they might need some of these same kinds of signals. I talked about neural architecture briefly in Section 2.8 of the last post, but mostly it's irrelevant to this series, and I won't talk about it beyond this unhelpfully-vague paragraph.

There might be other things in Category A that I'm not thinking of.

3.4.4 Category B: Everything else in the human Steering Subsystem (e.g. altruism-related drives)

I'll jump right into what I think is most important: **social instincts**, including various drives related to altruism, sympathy, love, guilt, remorse, status, jealousy, sense-of-fairness, etc. Key question: **How do I know that social instincts belong here in Category B, i.e. that they aren't one of the Category A things that are essential for general intelligence?**

Well, for one thing, look at high-functioning sociopaths. I've had the unfortunate experience of getting to know a couple of them very well in my day. They understood the world, and themselves, and language and math and science and technology, and they could make elaborate plans to successfully accomplish impressive feats. If there were an AI that could do everything that a high-functioning sociopath can do, we would *unhesitatingly* call it "AGI". Now, I think high-functioning sociopaths have *some* social instincts—they're more interested in manipulating people than manipulating toys—but their social instincts seem to be *very different* from those of a neurotypical person.

Then on top of that, we can consider people with autism, and people with schizophrenia, and SM (who is missing her amygdala and more-or-less lacks negative social emotions), and on and on. All these groups of people have "general intelligence", but their social instincts / drives are all quite different from each other's.[2]

All things considered, I find it very hard to believe that any aspect of social instincts is essential for general intelligence. I think it's at least open to question whether social instincts are even *helpful* for general intelligence!! For example, if you look at the world's most brilliant scientific minds, I'd guess that people with neurotypical social instincts are if anything slightly *under*represented.

One reason this matters is that, I claim, **social instincts underlie "the desire to behave ethically"**. Again, consider high-functioning sociopaths. They can *understand* honor and justice and ethics if they try—in the sense of correctly answering quiz questions about what is or isn't honorable etc.—they're just not *motivated* by it.[3]

If you think about it, it makes sense. Suppose I tell you "You really ought to put pebbles in your ears." You say "Why?" And I say "Because, y'know, your ears, they don't have any pebbles in them, but they really should." And again you say "Why?" …At some point, this conversation has to ground out at something that you find *inherently* motivating or demotivating, in and of itself. And I claim that social instincts—the various innate drives related to sense-of-fairness and sympathy and loyalty and so on—are ultimately providing the ground on which those intuitions stand.

(I'm not taking a stand on moral realism vs. moral relativism here—i.e., the question of whether there is a "fact of the matter" about what is ethical vs. unethical. Instead, I'm saying that *if* there's an agent that is completely lacking in any innate drives that might spur a desire to act ethically, then then we can't expect the agent to act ethically, no matter how intelligent and capable it is. Why would it? Granted, it might act ethically *as a means to an end*—e.g. to win allies—but that doesn't count. More discussion and intuition-pumps in my comment here.)

That's all I want to say about social instincts for now; I'll return to them in Post #13.

What else goes in Category B? Lots of things!! There's disgust, and aesthetics, and transcendence, and serenity, and awe, and hunger, and pain, and fear-of-spiders, etc.

3.4.5 Category C: Every other possibility (e.g. drive to increase my bank account balance)

When people make AGIs, they can put *whatever they want* into the reward function! This would be analogous to inventing new innate drives out of whole cloth. And these can be innate drives that are radically unlike anything in humans or animals.

Why might the future AGI programmers invent new-to-the-world innate drives? Because it's the obvious thing to do!! Go kidnap a random ML researcher from the halls of NeurIPS, drive them to an abandoned warehouse, and force them to make a bank-account-balance-increasing AI using reinforcement learning.[4] I bet you anything that, when you look at their source code, you're going to find a reward function that involves the bank account balance. You won't find anything like *that* among the genetically-hardwired circuitry in the human brainstem! It's a new-to-the-world innate drive.

Not only is “put in an innate drive for increasing the bank account balance” the obvious thing to do, but I think it would actually work! For a while! And then it would fail catastrophically! It would fail as soon as the AI became competent enough to find out-of-the-box strategies to increase the bank account balance—like borrowing money, hacking into the bank website, and so on. (Related: hilarious and terrifying list of historical examples of AIs finding unintended, out-of-the-box strategies for maximizing a reward. More on this in future posts.) In fact, this bank-account-balance example is one of the many, many possible drives that would plausibly lead to an AGI harboring a secret motivation to escape human control and kill everyone (see Post #1).

So these kinds of motivations are the worst: they’re dangling right in front of everyone’s faces, they’re the best way to get things done and publish papers and beat benchmarks if the AGI is not overly clever, and then when the AGI becomes competent enough, they lead to catastrophic accidents.

Maybe you’re thinking: “It’s *really obvious* that an AGI with an all-consuming innate drive to increase a certain bank account balance is an AGI that would try to escape human control, self-reproduce etc. Do you really believe that future AGI programmers would be *so reckless* as to put in something like that??”

Well, umm, yes. Yes, I do. But even setting that aside for the sake of argument, there’s a bigger problem: we don’t currently know how to code up *any innate drive whatsoever* such that the resulting AGI would definitely stay under control. Even the drives that *sound* benign are probably not, at least not in our current state of knowledge. Much more on this in later posts (especially #10).

To be sure, Category C is a very big tent. I would not be at all surprised if there exist Category C innate drives that would be *very good* for AGI safety! We just need to find them! I’ll be exploring this design space later in the series.

3.5 Brain-like AGI will by default have radically nonhuman (and dangerous) motivations

I mentioned this way back in the first post (Section 1.3.3), but now we have the explanation.

The previous subsection proposes three types of ingredients to put in a Steering Subsystem: (A) Those necessary to wind up with an AGI at all, (B) Everything else in humans, (C) Anything *not* in humans.

My claims are:

1. People want to make powerful AIs with state-of-the-art capabilities in challenging domains—they know that it’s good for publications, good for impressing their colleagues, getting jobs and promotions and grants, etc. I mean, just look at AI and ML today. Therefore, by default, I expect AGI researchers to race down the most direct path to AGI: reverse-engineering the Learning Subsystem, and combining it with Category-A drives.
2. Category B contains some drives that are plausibly useful for AGI safety: drives related to altruism, sympathy, generosity, humility, etc. Unfortunately, we don’t currently know how any of those drives are implemented in the brain. And figuring that out is unnecessary for building AGIs. So by default, I think we should expect AGI researchers to ignore Category B until they have AGIs up and running, and *only then* start scrambling to figure out how to build altruism drives etc. And they might outright fail—it’s totally possible that the corresponding brainstem & hypothalamus circuitry is a frightfully complicated mess, and we only have so much time between “AGIs are up and running” and “someone accidentally makes an out-of-control AGI that kills everyone” (see Post #1).
3. There are things in Category C like *“A low-level innate drive to increase a particular bank account balance”* that are immediately obvious to everyone, and easy to implement, and

> will work well at accomplishing the programmers' goals *while their janky proto-AGIs are not yet very capable*. Therefore, by default, I expect future researchers to use these kinds of "obvious" (but dangerous and radically-nonhuman) drives as they work towards developing AGI. And as discussed above (and more in later posts), even if the researchers start trying in good faith to give their AGI an innate drive for being helpful / docile / whatever, they might find that they don't know how to do so.

In sum, if researchers travel down the most easy and natural path—the path that looks like the AI and neuroscience R&D community continuing to behave in ways that they behave right now—we will wind up being able to make AGIs that do impressive things that their programmers want, for a while, but are driven by radically alien motivation systems that are fundamentally unconcerned with human welfare, and these AGIs will try to escape human control as soon as they are capable enough to do so.

Let's try to change that! In particular, if we can figure out *in advance* how to write code that builds an innate drive for altruism / helpfulness / docility / whatever, that would be a huge help. This will be a major theme of this series. But don't expect final answers. It's an unsolved problem; there's still a lot of work to do.

3.6 Response to Jeff Hawkins's argument against AGI accident risk

Jeff Hawkins has a recent book *A Thousand Brains*. I wrote a more detailed book review here. Jeff Hawkins is a strong advocate of a two-subsystems perspective very similar to mine. No coincidence—his writings helped push me in that direction!

To Hawkins's great credit, he takes ownership of the idea that his neuroscience / AI work is pushing down a path (of unknown length) towards AGI, and he has tried to think carefully about the consequences of that larger project—as opposed to the more typical perspective of declaring AGI to be someone else's problem.

So, I'm delighted that Hawkins devotes a large section of his book to an argument about AGI catastrophic risk. But his argument is *against* AGI catastrophic risk!! What's the deal? How do he and I, starting from a similar two-subsystems perspective, wind up with diametrically opposite conclusions?

Hawkins makes many arguments, and again I addressed them more comprehensively in my book review. But here I want to emphasize two of the biggest issues that bear on this post.

Here's my paraphrase of a particular Hawkins argument. (I'm translating it into the terminology I'm using in this series, e.g. he says "old brain" where I say "Steering Subsystem". And maybe I'm being a bit mean. You can read the book and judge for yourself whether this is fair.)

1. The Learning Subsystem (neocortex etc.) *by itself* has no goals or motivations. It won't do anything. It certainly won't do anything dangerous. It's like a map sitting on a table.
2. Insofar as humans have problematic drives (greed, self-preservation, etc.), they come from the Steering Subsystem (brainstem etc.).
3. The thing that I, Jeff Hawkins, am proposing, and doing, is trying to reverse-engineer the Learning Subsystem, not the Steering Subsystem. So what the heck is everyone so worried about?
4. …
5. …
6. Oh hey, on a *completely* unrelated note, we will eventually make future AGIs, and these will have not only a Learning Subsystem, but also a Steering Subsystem attached to it. I'm not going to talk about how we'll design the Steering Subsystem. It's not really something that I think about much.

Each of these points *in isolation* seems reasonable enough. But when you put them together, there's a gaping hole! Who cares if a neocortex *by itself* is safe? A neocortex *by itself* was never the plan! The question we need to ask is whether an AGI consisting of *both* subsystems attached together will be safe. And that depends crucially on how we build the Steering Subsystem. Hawkins isn't interested in that topic. But I am! Read on in the series for much more on this. Post #10 in particular will dive into why it's a heck of a lot harder than it sounds to build a Steering Subsystem that steers the AGI into doing some particular thing that we intend for it to do, without also incidentally instilling dangerous antisocial motivations that we never intended it to have.

One more (related) issue that I didn't mention in my earlier book review: I think that Hawkins is partly driven by an intuition that I argued against in (Brainstem, Neocortex) ≠ (Base Motivations, Honorable Motivations) (and more on that topic coming up in Post #6): a tendency to inappropriately locate ego-syntonic motivations like "unraveling the secrets of the universe" in the neocortex (Learning Subsystem), and ego-dystonic motivations like hunger and sex drive in the brainstem (Steering Subsystem). I claim that the correct answer is that *all* motivations come ultimately from the Steering Subsystem, no exceptions. This will hopefully be obvious if you keep reading this series.

In fact, my claim is even implied by the better parts of Hawkins's own book! For example:

- Hawkins in Chapter 10: "The neocortex learns a model of the world, which by itself has no goals or values."
- Hawkins in Chapter 16: " 'We'—the intelligent model of ourselves residing in the neocortex—are trapped. We are trapped in a body that…is largely under the control of an ignorant brute, the old brain. We can use intelligence to imagine a better future…. But the old brain could ruin everything…"

To spell out the contradiction: if "we" = the neocortex's model, and the neocortex's model has no goals or values whatsoever, then "we" certainly would not be aspiring to a better future and hatching plots to undermine the brainstem.

3.7 Timelines-to-brain-like-AGI part 2 of 3: how hard will it be to reverse-engineer the Steering Subsystem well enough for AGI?

(Reminder: Timelines Part 1 of 3 was Section 2.8 of the previous post .)

Above (Section 3.4.3), I discussed "Category A", the minimal set of ingredients to build an AGI-capable Steering Subsystem (not necessarily *safe*, just *capable*).

I don't *really* know what is in this set. I suggested that we'd probably need some kind of curiosity drive, and maybe some drive to pay attention to human language and other human activities, and maybe some signals that go along with and help establish the Learning Subsystem's neural network architecture.

If that's right, well, this doesn't strike me as too hard! Certainly it's a *heck* of a lot easier than reverse-engineering everything in the human hypothalamus and brainstem! Keep in mind that there is a substantial literature on curiosity in both ML (1, 2) and psychology. "A drive to pay attention to human language" requires nothing more than a classifier that says (with reasonable accuracy, it doesn't have to be perfect) whether any given audio input is or isn't human language; that's *trivial* with today's tools, if it's not already on GitHub.

I think we should be open to the possibility that it just isn't that hard to build a Steering Subsystem that (together with a reverse-engineered Learning Subsystem, see Section 2.8 of the previous post) can develop into an AGI after training. Maybe it's not decades of R&D; maybe it's not even years of R&D! Maybe a competent researcher will nail it after just a couple tries. On the other hand—maybe not! Maybe it *is* super hard! I think it's very difficult to predict how long it would take, from our current vantage point, and that we should remain uncertain.

3.8 Timelines-to-brain-like-AGI part 3 of 3: scaling, debugging, training, etc.

Having a fully-specified, AGI-capable algorithm isn't the end of the story; you still need to implement the algorithm, iterate on it, hardware-accelerate and parallelize it, work out the kinks, run trainings, etc. We shouldn't *ignore* that part, but we shouldn't overstate it either. I won't get into this here, because I recently wrote a whole separate blog post about it:

Brain-inspired AGI and the "lifetime anchor"

The upshot of that post is: I think all that stuff could absolutely get done in <10 years. Maybe even <5. Or it could take longer. I think we should be very uncertain.

Thus concludes my timeline-to-brain-like-AGI discussion, which again is not my main focus in this series. You can read my three timelines sections (2.8, 3.7, and this one), agree or disagree, and come to your own conclusions.

3.9 Timelines-to-brain-like-AGI *encore*: How should I *feel* about a probabilistic timeline?

My "timelines" discussion (Sections 2.8, 3.7, 3.8) has been about the *forecasting* question "what probability distribution should I assign to when AGI will arrive (if ever)?"

Semi-independent of that question is a kind of *attitude* question: "How should I *feel* about that probability distribution?"

For example, there can be two people who *both* agree with (just an example) "35% chance of AGI by 2042". But their *attitudes* may be wildly different:

- One of the two people rolls their eyes, laughs, and says: "See, I told you! AGI probably isn't coming for *decades*!"
- The other person widens their eyes, drops their jaw, and says "Oh. My. God. Excuse me for a moment while I rethink everything about my life."

There are a lot of factors underlying these different attitudes towards the same belief about the world. First, some factors are kinda more questions of psychology rather than questions of fact:

- *"What attitude better fits my self-image and psychology?"*—ooh, yikes, this one cuts deep into our psyches. People who think of themselves as cool-headed serious skeptical dignified grounded scientists may feel irresistibly drawn to the belief that AGI isn't a big deal. People who think of themselves as pioneering radical transhumanist technologists may equally feel irresistibly drawn to the opposite belief that AGI will radically change *everything*. I bring this up so that you can meditate on your own biases. Oh, who am I kidding; realistically, I just handed you a nice way to smugly mock and dismiss anyone who disagrees with you. (You're welcome!) For my part, I claim some immunity to being dismissed-via-psychoanalysis: When I first came to believe that AGI is a very big deal, I *totally* self-identified as a cool-headed serious skeptical dignified grounded middle-aged scientist, with no interest in, nor connection to, science fiction or transhumanism or the tech industry or AI or silicon valley, etc. Take that! Ha! But really, this is a stupid game to play: dismissing people's beliefs by psychoanalyzing them for hidden motives has always been a terrible idea. It's too easy. Right or wrong, you can *always* find a good reason to smugly question the motives of anyone you disagree with. It's just a cheap trick to avoid the hard work of figuring out whether they might actually be correct. Also on the general

topic of psychology: taking our possible AGI future seriously (as seriously as I think is warranted) can be, well, kinda wrenching! It was hard enough getting used to the idea that Climate Change is really happening, right?? See this post for more on that.

- *How should I think about possible-but-uncertain future events?* I suggest reading this Scott Alexander post. Or if you prefer the meme version:

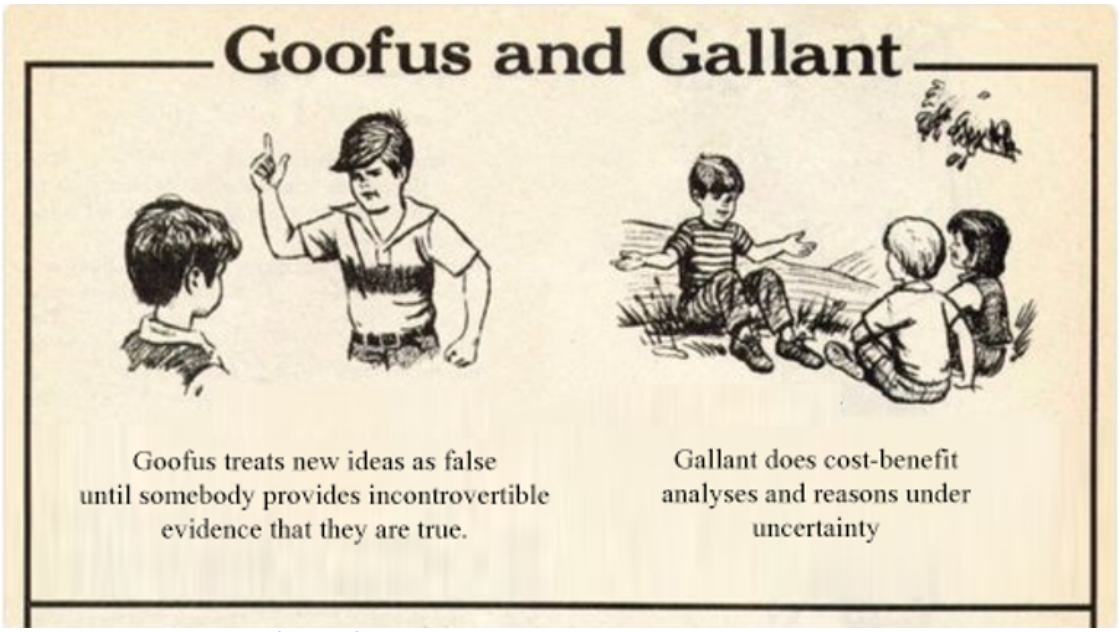
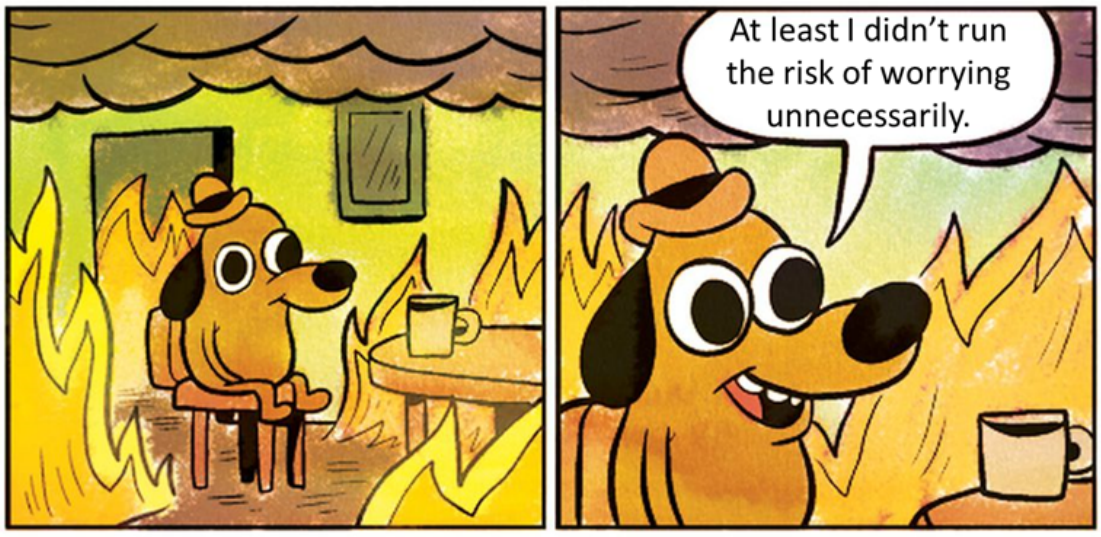


Image source: Scott Alexander.

Relatedly, there's a kind of feeling expressed by the famous "Seeing the Smoke" essay, and this meme here:



Loosely based on a @Linch meme, if I recall correctly.

To spell it out, the *right* idea is to weigh risks and benefits and probabilities of over-preparing vs. under-preparing for an uncertain future risk. The *wrong* idea is to add an extra entry into that ledger—"the risk of looking foolish in front of my friends by over-preparing for something weird that winds up not being a big deal"—and treat that one entry as *overwhelmingly more important than everything else on the list*, and then it follows that we shouldn't try to mitigate a possible future catastrophe until we're >99.9% confident that the catastrophe will definitely happen, in a

kind of insane bizarro-world reversal of Pascal's Wager. Luckily, this is increasingly a moot point; your friends are less and less likely to think you're weird, because AGI safety has gotten much more mainstream in recent years—thanks especially to outreach and pedagogy by Stuart Russell, Brian Christian, Rob Miles, and many others. You can help that process along by sharing this post series! ;-)

Putting those aside, other reasons for different attitudes towards AGI timelines are more substantive, particularly the questions:

- *How much will AGI transform the world?* For my part, I'm way the heck over on the "lots" end of the spectrum. I endorse the Eliezer Yudkowsky quote: "Asking about the effect of [superhuman AGI] on [unemployment] is like asking how US-Chinese trade patterns would be affected by the Moon crashing into the Earth. There would indeed be effects, but you'd be missing the point." For a more sober discussion, try Holden Karnofsky's Digital People Would Be An Even Bigger Deal, and maybe also This Can't Go On as background, and what the heck, the whole rest of that series too. Also see here for some numbers suggesting that brain-like AGI will probably *not* require so many computer chips or so much electricity that it can't be widely used.
- *How much do we need to do, to prepare for AGI?* See Post #1, Section 1.7 for my argument that we're way behind schedule, and later in this series I'll be discussing the many still-unsolved problems.

1. ^

   Well, maybe *some* people expect that there's a one-to-one correspondence between English-language abstract concepts like "sadness" and corresponding innate reactions. If you read the book *How Emotions Are Made*, Lisa Feldman Barrett spends hundreds of pages belaboring this point. She must have been responding to *somebody*, right? I mean, it feels to me like an absurd straw-man to say "Each and every situation that a native English speaker would describe as 'sadness' corresponds to the exact same innate reaction with the exact same facial expression." I'd be surprised if even Paul Ekman (whom Barrett was supposedly rebutting) actually believes that, but I dunno.

2. ^

   I wouldn't suggest that the Steering Subsystem circuitry underlying social instincts is built in a fundamentally different way in these different groups—that would be evolutionarily implausible. Rather, I think there are lots of adjustable parameters on how strong the different drives are, and they can be set to wildly different values, including the possibility that a drive is set to be so weak as to be effectively absent. See my speculation on autism and psychopathy here.

3. ^

   See Jon Ronson's *The Psychopath Test* for a fun discussion of attempts to teach empathy to psychopaths. The students merely wound up better able to *fake* empathy in order to manipulate people. Quote from one person who taught such a class: "I guess we had inadvertently created a finishing school for them."

4. ^

   I suppose I could have *hired* an ML researcher instead. But who could afford the salary?

[Intro to brain-like-AGI safety] 4. The “short-term predictor”

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

4.1 Post summary / Table of contents

Part of the “Intro to brain-like-AGI safety” post series .

The previous two posts (#2, #3) presented a big picture of the brain, consisting of a Steering Subsystem (brainstem and hypothalamus) and Learning Subsystem (everything else), with the latter “learning from scratch” in a particular sense defined in Post #2.

I suggested that our explicit goals (e.g. “I want to be an astronaut!”) emerge from an interaction between these two subsystems, and that understanding that interaction is critical if we want to assess how to sculpt the motivations of a brain-like AGI, so that it winds up trying to do things that we want it to be trying to do, and thus avoid the kinds of catastrophic accidents I discussed in Post #1.

These next three posts (#4–#6) are working our way up to that story. This post provides an ingredient that we’ll need: “the short-term predictor”.

Short-term prediction is *one* of the things the Learning Subsystem does—I’ll talk about others in future posts. A short-term predictor has a supervisory signal (a.k.a. “ground truth”) from somewhere, and then uses a learning algorithm to build a predictive model that anticipates that signal a short time (e.g. a fraction of a second) in the future.

This post will be a general discussion of how short-term predictors work and why they’re important. They will turn out to be a key building block of motivation and reinforcement learning, as we’ll see in the subsequent two posts.

Teaser for the next couple posts: The next post (#5) will discuss how a certain kind of closed-loop circuit wrapped around a short-term predictor turns it into a “long-term predictor”, which has connections to the temporal difference (TD) learning algorithm. I will argue that the brain has a large number of these long-term predictors, built out of telencephalon-brainstem loops, one of which is akin to the “critic” part of actor-critic reinforcement learning. The “actor” part is the subject of Post #6.

Table of contents:

- Section 4.2 gives a motivating example, of flinching just before getting hit in the face. This can be formulated as a supervised learning problem, in the sense that there is a ground-truth signal to learn from. (If you just got hit in the face, then you should have flinched!) The resulting circuit is what I call a “short-term predictor”.
- Section 4.3 defines terminology: “context signals”, “output signals”, and “supervisory signals”. (In ML terminology, these correspond respectively to “trained model inputs”, “trained model outputs”, and “labels”.)
- Section 4.4 offers a sketch of an extremely simple short-term predictor that could be built out of biological neurons, just so you can have something concrete in mind.
- Section 4.5 discusses the benefits of short-term predictors compared to alternative approaches including (in the flinching example) a hardwired circuit for deciding when to flinch, or a reinforcement learning (RL) agent that is rewarded for appropriate flinching. For the latter, a short-term predictor can learn faster than an RL agent because it gets an error gradient “for free” each query—or in simpler terms, when it

screws up, it gets some indication of what it did wrong, e.g. whether the error is an overshoot vs. undershoot.
- Sections 4.6-4.8 cover various examples of short-term predictors in the human brain. None of these are especially important for AGI safety—the *really* important one is the topic of the next post—but they come up sufficiently often that they warrant a brief discussion:
  - Section 4.6 covers the cerebellum, with my theory that it's a collection of ≈300,000 short-term predictors, used to (in effect) reduce the latency on ≈300,000 signals traveling around the brain and body.
  - Section 4.7 covers predictive learning of sensory inputs in the cortex—i.e., you're constantly predicting what you're about to see, hear, feel, etc., and the corresponding prediction errors are used to update your internal models.
  - Section 4.8 briefly covers a few other neat random things that short-term predictor circuits can do for an animal.

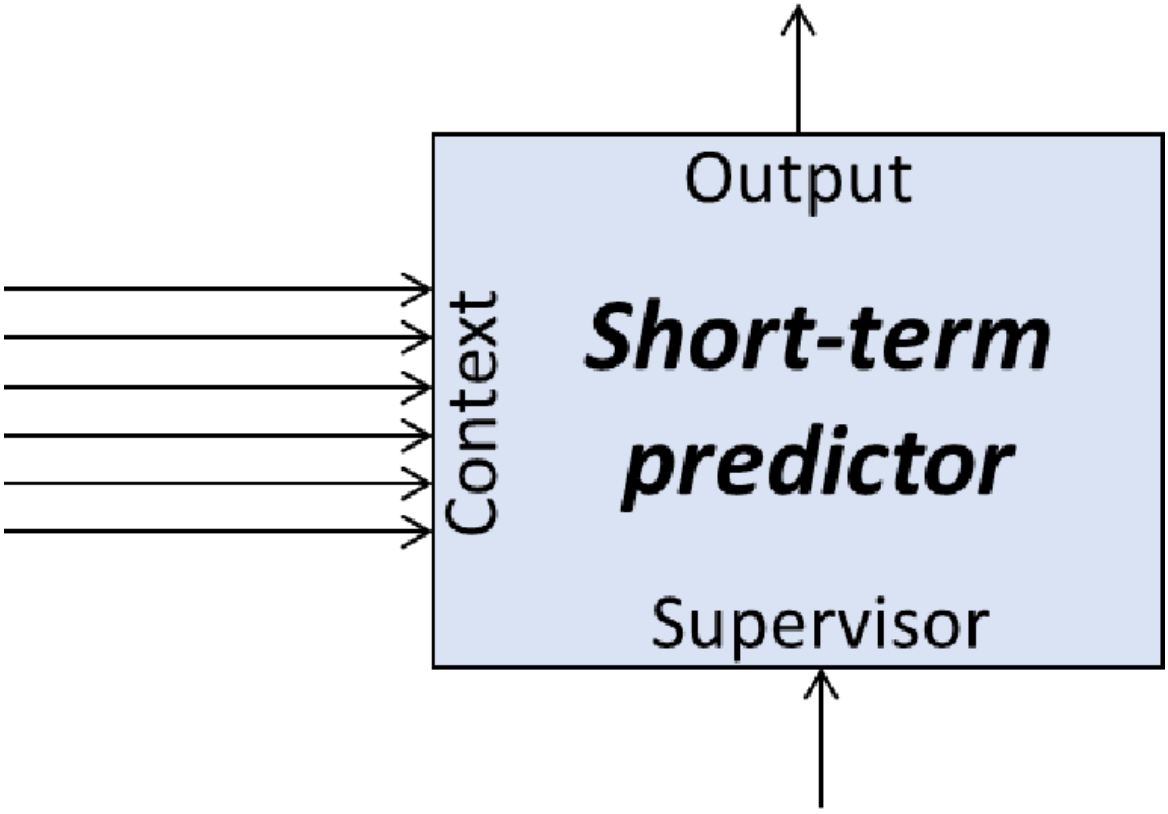
4.2 Motivating example: flinching before getting hit in the face

Suppose you have a job or hobby in which there's a particular, recognizable sensory cue (e.g. someone yelling "FORE!!!" in golf), and then half a second after that cue you very often get whacked in the face. Your brain is going to *learn* to (involuntarily) flinch in response to the cue. There's a learning algorithm inside your brain, commanding these flinches; it presumably evolved to protect your face. That learning algorithm is what I want to talk about in this post.

I'm calling it a "short-term predictor". It's a "predictor" because the goal of the algorithm is to predict something in advance (i.e., an upcoming whack in the face). It's "short-term" because we only need to predict what will happen a fraction of a second into the future. It's more specifically a type of supervised learning algorithm, because there is a "ground truth" signal indicating what the prediction output *should* have been in hindsight.

4.3 Terminology: Context, Output, Supervisor

Our "short-term predictor" has three ingredients in its "API" ("application programming interface"—i.e., the channels through which other parts of the brain interact with the "short-term predictor" module):



- An **output signal** is the algorithm's prediction.
  - In our example above, this would be a signal that triggers a flinch reaction.
- A **supervisory signal** provides "ground truth" (in hindsight) about what the algorithm's output *should have* been.
  - In our example above, this would be a signal that indicates that I just got whacked in the face (and therefore, implicitly, I *should have* flinched).
  - In ML terminology, "supervisory signals" are often called "labels".
  - In the actual implementation, the supervisor-type input to the short-term predictor does not *have* to be the ground truth. It could also be an error signal, or a negative error signal, etc. From my perspective, this is an unimportant low-level implementation detail.
- **Context signals** carry information about what's going on.
  - In our example above, this might be a random assortment of signals (corresponding to latent variables) coming from the visual cortex and auditory cortex. With luck, some of those signals might carry predictively-useful information: maybe one signal conveys the fact that I am on a golf course, and another signal conveys the fact that someone near me just yelled "FORE!".
  - In ML terminology, "context signals" would instead be called "inputs to the trained model".

The context signals don't *all* have to be relevant to the prediction task. We can just throw a whole bunch of crap in there, and the learning algorithm will automatically go searching for the context data that are useful for the prediction task, and ignore everything else.

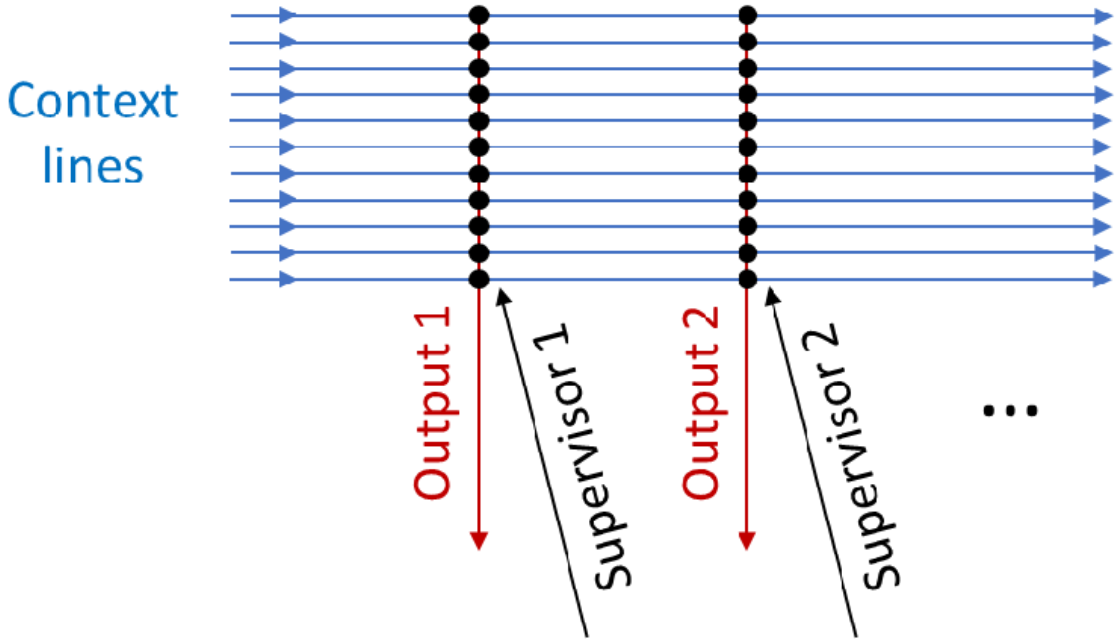
4.4 Extremely simplified toy example of how this could work in biological neurons

How might a short-term predictor work at a low level?

Well, suppose we want an output signal that precedes the supervisor signal by 0.3 seconds—as above, for example, maybe we want to learn to flinch *before* getting hit. We grab a bunch of context data that might be relevant—for example, neurons carrying partially-processed visual information. We track which of those context lines is disproportionately likely to fire 0.3 seconds before the supervisor does. Then we wire up those context lines to the output.

And we're done! Easy peasy.

In biology, this would look something like synaptic plasticity with a "three-factor learning rule"—i.e., the synapse gets stronger or weaker as a function of the activity of three different neurons (context, supervisor, output), and their relative timings.



Black dots indicate adjustable-strength synapses

To be clear, a short-term predictor can be *much, much* more complicated than this. Making it more complicated can give better performance. To pick a fun example that I just learned about the other day, apparently the short-term predictors in the cerebellum (Section 4.6 below) have neurons that can somehow *store an adjustable time-delay parameter within the neuron itself* (!!) (ref—it came up on this podcast). Other possible bells and whistles include pattern separation (Post #2, Section 2.5.4), and training multiple outputs with the same supervisor and pooling them (ref), or better yet training multiple outputs with the same supervisor but with different hyperparameters, in order to get a probability distribution (original paper, further discussion here), and so on.

So this subsection is an oversimplification. But I won't apologize. I think these kinds of grossly-oversimplified toy models are important to talk about and keep in mind. From a *conceptual* perspective, we get to feel like there's probably no deep mystery hidden behind the curtain. From an *evolutionary* perspective, we get to feel like there's a plausible story of how early animals can start with a very simple (but still useful) circuit, and the circuit can get gradually more complicated over many generations. So get used to it—many more grossly-oversimplified toy models are coming up in future posts!

4.5 Comparison to other algorithmic approaches

4.5.1 “Short-term predictor” versus a hardwired circuit

Let’s go back to the example above: flinching before getting whacked in the face. I suggested that a good way to decide when to flinch is with a “short-term predictor” learning algorithm. Here’s an alternative: we can *hardwire* a circuit that decides when to flinch. For example, if there’s a blob in the field-of-view whose size is rapidly increasing, then it’s probably a good time to flinch. A detector like that could plausibly be hardwired into the brain.

How do those two solutions compare? Which is better? Answer: no need to decide! They’re complementary. We can have both. But still, it’s pedagogically helpful to spell out their comparative advantages and disadvantages.

The main (only?) advantage of the hardwired flinching system is that it works from birth. Ideally, you wouldn’t get whacked in the face even once. By contrast, the short-term predictor is a learning algorithm, and thus generally needs to “learn things the hard way”.

In the other direction, the short-term predictor has two powerful advantages over the hardwired solution—one obvious, one not-so-obvious.

The obvious advantage is that a short-term predictor is powered by within-lifetime learning, not evolution, and therefore can learn cues for flinching that were rarely or never present in previous generations. If I tend to bonk my head whenever I walk into a certain cave, I’ll learn to flinch. There’s no chance that my ancestors evolved a reflex to flinch at *this* particular part of *this* particular cave. My ancestors might have never been to this cave. The cave might not have existed until last week!

The less obvious, but very important, advantage is that a short-term predictor can learn patterns that involve learned-from-scratch patterns (Post #2), whereas a hardwired flinching system can’t. The argument here is the same as Section 3.2.1 of the previous post: the genome cannot know *exactly* which neurons (if any) will store any particular learned-from-scratch pattern, and therefore cannot hardwire a connection to them.

The ability to leverage learned-from-scratch patterns is a big benefit. For example, there may well be good cues for flinching that depend on learned-from-scratch semantic patterns (e.g. the knowledge “I am playing golf right now”), learned-from-scratch visual patterns (e.g. the visual appearance of a person swinging a golf club) or learned-from-scratch location tags (e.g. “this particular room, which has a low ceiling”), and so on.

4.5.2 “Short-term predictor” vs an RL agent: Faster learning thanks to error gradients

The short-term prediction circuit is a special case of *supervised learning*.

Supervised learning is when you have a learning algorithm receiving a ground-truth signal like this:

“Hey learning algorithm: you messed up—you should have done thus-and-such instead.”

Compare that to reinforcement learning (RL), where the learning algorithm gets a *much less helpful* ground-truth signal:

"Hey learning algorithm: you messed up."

(a.k.a negative reward). Obviously, you can learn much faster with supervised learning than with reinforcement learning. The supervisory signals, at least in principle, tell you exactly what parameter settings to change and how, if you want to do better next time you're in a similar situation. Reinforcement learning doesn't; instead you need trial-and-error.

In technical ML terms, supervised learning provides a full error gradient "for free" on each query, whereas reinforcement learning does not.

Evolution can't *always* use supervised learning. For example, if you're a professional mathematician trying to prove a theorem, and your latest proof attempt didn't work, there is no "ground truth" signal that says what to do differently next time—not in your brain, not out there in the world. Sorry! You're in a very-high-dimensional space of possible things to do, with no real guideposts. At some level, trial-and-error is your only option. Tough luck.

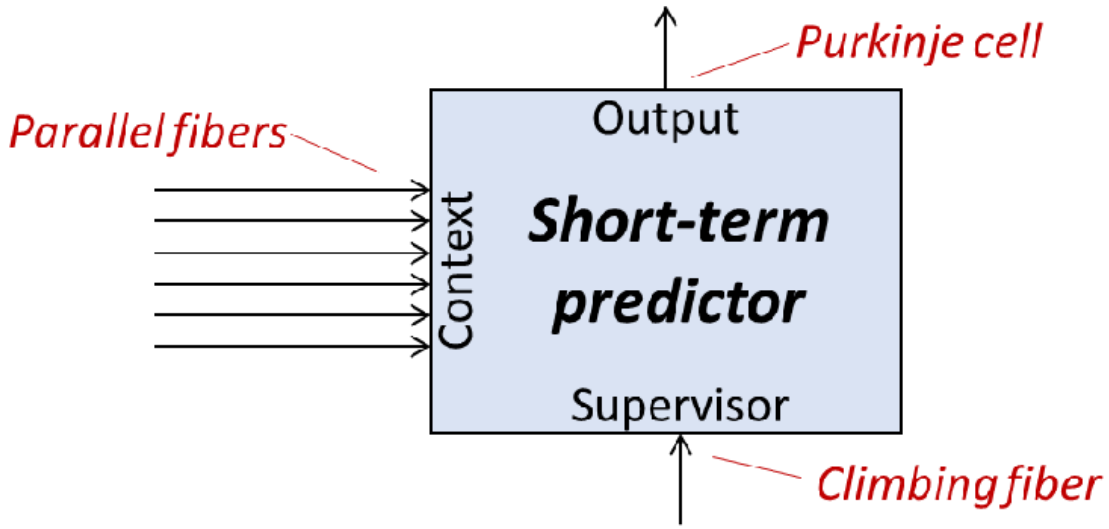
But evolution *can sometimes* use supervised learning, as in the examples in this post. And my point is: if it *can*, it probably *will*.

4.6 "Short-term predictor" example #1: The cerebellum

I'll jump right into what I think the cerebellum is for, and then I'll talk about how my theory relates to other proposals in the literature.

4.6.1 My theory of the cerebellum

My claim is that the cerebellum is housing lots short-term prediction circuits.



Relation of cerebellum neuroanatomy (red) with our diagram from above. As usual (see above), I'm leaving out lots of bells and whistles that make the short-term predictor more accurate, like [there's an extra layer I'm

not showing, plus pattern separation (Post #2, Section 2.5.4), etc.

How many short-term predictors? My best guess is: around 300,000 of them.[1]

What on earth?? Why oh why does your brain need 300,000 short-term predictors?

I have an opinion! I think the cerebellum sits there, watching lots of signals in the brain, and *it learns to preemptively send those same signals itself*.

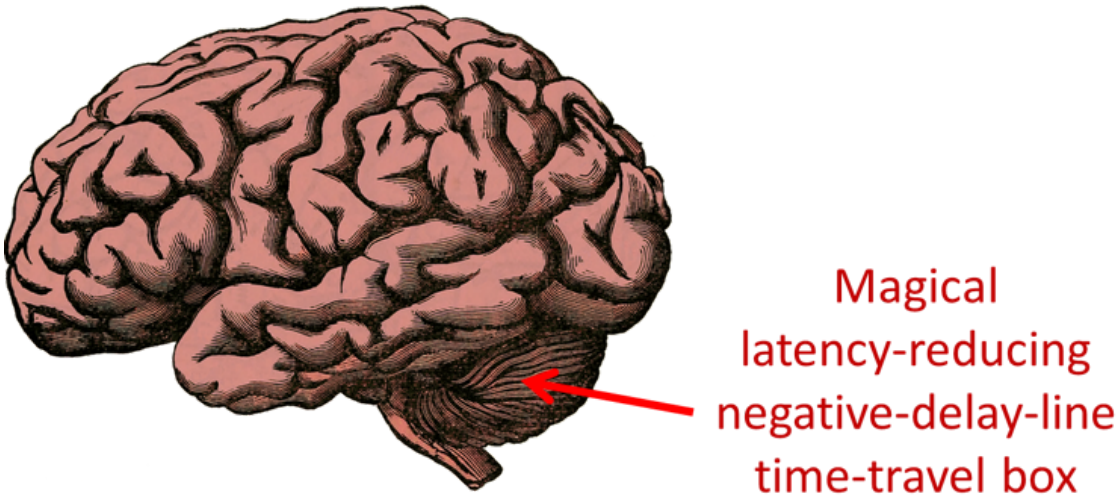
That's it. That's my whole theory of the cerebellum.

In other words, the cerebellum might discover the rule "Given the current context information, I predict that cortical output neuron #187238 is going to fire in 0.3 seconds". Then the cerebellum goes ahead and sends a signal *right now*, to the same place. Or in the opposite direction, the cerebellum might discover the rule "Given the current context information, I predict that proprioceptive nerve #218502 is going to fire in 0.3 seconds". Again, the cerebellum goes ahead and sends a signal *right now*, to the same place.

Some vaguely-analogous concepts:

- When the cerebellum is predicting-and-preempting the telencephalon, we can think of it as vaguely akin to "memoization" in software engineering, or "knowledge distillation" in machine learning, or this recent paper proposing (so-called) "neural surrogates".
- When the cerebellum is predicting-and-preempting peripheral nerves, we can think of it as building a bunch of predictive models of the body, each narrowly-tailored to predict a different peripheral nerve signal. Then when the telencephalon is doing motor control, and needs peripheral feedback signals, it can use those predictive models as feedback, instead of the real thing.

Basically, I think the brain has these issues where the *throughput (a.k.a. bandwidth)* of a subsystem is adequate, but its *latency* is too high. In the peripheral nerve case, the latency is high because the signals need to travel a great distance. In the telencephalon case, the latency is high because the signals need to travel a shorter-but-still-substantial distance, and moreover need to pass through multiple sequential processing steps. In any case, the cerebellum can magically reduce the latency, at the cost of occasional errors. The cerebellum sits in the middle of the action, always saying to itself "what signal is about to appear here?", and then it preemptively sends it. And then a fraction of a second later, it sees whether its prediction was correct, and updates its models if it wasn't. It's like a little magical time-travel box—a delay line whose delay is negative.



And now we have our answer: why do we need ≈300,000 short-term predictors? Because there are lots of peripheral nerves, and there are lots of telencephalon output lines, and maybe other things too. And a great many of those signals can benefit from being predicted-and-preempted! Heck, if I understand correctly, the cerebellum can even predict-and-preempt a signal *that goes from the telencephalon to a different part of the telencephalon*!

That's my theory. I haven't run simulations or anything; it's just an idea. See here and here for two examples in which I've used this model to try to understand observations in neuroscience and psychology. Everything else I know about the cerebellum—neuroanatomy, how it's connected to other parts of the brain, lesion and imaging studies, etc.—all seem to fit this theory really well, as far as I can tell. But really, this little section is almost the sum total of what I know about this topic.

4.6.2 How my cerebellum theory relates to others in the literature

(I'm not an expert here and am open to correction.)

I think it's widely agreed that the cerebellum is involved in supervised learning. I believe that idea is called the Marr-Albus-Ito model, cf. Marr 1969 or Albus 1971, or the fun Brains Explained YouTube channel.

Recall from above that a short-term predictor is an example of a supervised learning algorithm, but supervised learning is a broader category. So the supervised learning part is *not* a distinguishing feature of my proposal above, and in particular that diagram above (with cerebellum neuroanatomy in red) is compatible with the usual Marr-Albus-Ito story. Instead, the distinguishing aspect of my theory concerns what the ground truth signals are (or what the error signals are—which amounts to the same thing).

I mentioned in Post #2 that when I see a within-lifetime learning algorithm, my immediate question is: "What's the ground truth that it's learning from?" I also mentioned that usually, when I go looking for an answer in the literature, I wind up feeling confused and unsatisfied. The cerebellum literature is a perfect example.

For example, I often hear something to the effect of "cerebellar synapses are updated when there's a motor error". But who's to say what constitutes a motor error?

- If you're trying to walk to school, then slipping on a banana peel is a motor error.
- If you're trying to slip on a banana peel, then slipping on a banana peel is bang-on!

How is the cerebellum supposed to know? I don't get it.

I've read a number of computational theories of the cerebellum. They tend to be *way* more complicated than mine. And they *still* leave me feeling like I don't understand where the ground truth is coming from. To be clear, I haven't carefully read every paper and it remains possible that I'm missing something.

*(**Update July 2022:** Hooray, I found this 2006 paper by Harri Valpola which suggests essentially the same cerebellum model as mine above. Check it out for helpful discussion including further references to the literature.)*

Well, whatever. It doesn't really impact this series. As I mentioned earlier, you can be a functioning adult able to live independently, hold down a job, etc., without a cerebellum at all. So if I'm totally wrong about the cerebellum, it shouldn't really impact the big picture.

4.7 "Short-term predictor" example #2: Predictive learning of sensory inputs in the cortex

Your cortex has a rich generative model of the world, including yourself. Every fraction of a second, your brain uses that model to predict incoming sensory inputs (sight, sound, touch, proprioception, interoception, etc.), and when its predictions are wrong, the model is updated on the error. Thus, for example, you can open your closet door, and know *immediately* that somebody oiled the hinge. You were predicting that it would sound and feel a certain way, and the prediction was falsified.

In my view, predictive learning of sensory inputs is the jumbo jet engine bringing information from the world into our cortical world-model. I endorse the Yann LeCun quote: "If intelligence is a cake, the bulk of the cake is [predictive learning of sensory inputs], the icing on the cake is [other types of] supervised learning, and the cherry on the cake is reinforcement learning." The sheer number of bits of data we get from predictive learning of sensory inputs swamps everything else.

Predictive learning of sensory inputs—in the specific sense I'm using it here—is not a grand unified theory of cognition. The big problem occurs when it collides with "decisions" (what muscles to move, what to pay attention to, etc.). Consider the following: I can predict that I'll sing, and then I sing, and my prediction was correct. Or I can predict that I'll dance, and then I dance, and then *that* prediction was correct. Thus, predictive learning is at a loss; it can't help me do the right thing here. That's why we *also* need the Steering Subsystem (Post #3) to send supervisory signals and RL reward signals. Those signals can promote good decisions in a way that predictive learning of sensory inputs cannot.

Nevertheless, predictive learning of sensory inputs is a very big deal for the brain, and there's a lot to be said about it. However, I've come to see it as one of many topics that seems very directly important for *building* a brain-like AGI, but only slightly relevant for brain-like-AGI *safety*. So I'll mention it from time to time, but if you're looking for gory details, you're on your own.

4.8 Other example applications of "short-term predictors"

These also won't be important for this series, so I won't say much about them, but just for fun, here are three more random things that I think Evolution can do with a short-term predictor circuit.

- Filtering—for example, my brain can make a short-term predictor of my audio input stream, with the constraint that its context inputs *only* carry information about my own jaw motion and my own vocal cord activity. The predictor should wind up with a model of purely the self-generated contribution to my audio input stream. That's very useful because my brain can *subtract it off*, leaving only externally-generated sounds.
- Input data compression—this is kinda a more extreme version of filtering. Instead of merely filtering out information that's predictable from self-generated activity, we filter out information that's predictable from *any information whatsoever that we already know*. By the way, this is how I'm tentatively thinking about the dorsal cochlear nucleus, a little structure in the auditory input processing chain that looks *suspiciously* like the cerebellum. See here. Warning: It's possible that this idea makes no sense; I go back and forth.
- Novelty detection—see discussion here.

1. ^

   There are 15 million Purkinje cells (ref), but this paper says that one predictor consists of "a handful of" Purkinje cells with a single supervisor and a single (pooled) output. What does "handful" mean? The paper says "around 50". Well, 50 in mice. I can't immediately find the corresponding number for humans. I'm assuming it's still 50, but that's just a guess. Anyway, that's how I wound up guessing that there are 300,000 predictors

[Intro to brain-like-AGI safety] 5. The “long-term predictor”, and TD learning

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

5.1 Post summary / Table of contents

Part of the “Intro to brain-like-AGI safety” post series .

In the previous post, I discussed the “short-term predictor”—a circuit which, thanks to a learning algorithm, emits an output that predicts a ground-truth supervisory signal arriving a short time (e.g. a fraction of a second) later.

In this post, I propose that we can take a short-term predictor, wrap it up into a closed loop involving a bit more circuitry, and we wind up with a new module that I call a “long-term predictor”. Just like it sounds, this circuit can make longer-term predictions, e.g. “I’m likely to eat in the next 10 minutes”. This circuit is closely related to Temporal Difference (TD) learning, as we’ll see.

I will argue that there are a large collection of side-by-side long-term predictors in the brain, each comprising a short-term predictor in the telencephalon (involving specific areas such as ventral striatum, medial prefrontal cortex, and amygdala) that loops down to the Steering Subsystem (hypothalamus and brainstem) and then back via a dopamine neuron. These long-term predictors make predictions about biologically-relevant inputs and outputs—for example, one long-term predictor might predict whether I’ll feel pain in my arm, another whether I’ll get goosebumps, another whether I’ll release cortisol, another whether I’ll eat, and so on. Moreover, one of these long-term predictors is essentially a value function for reinforcement learning.

All these predictors will play a major role in motivation—a story which I will finish in the next post.

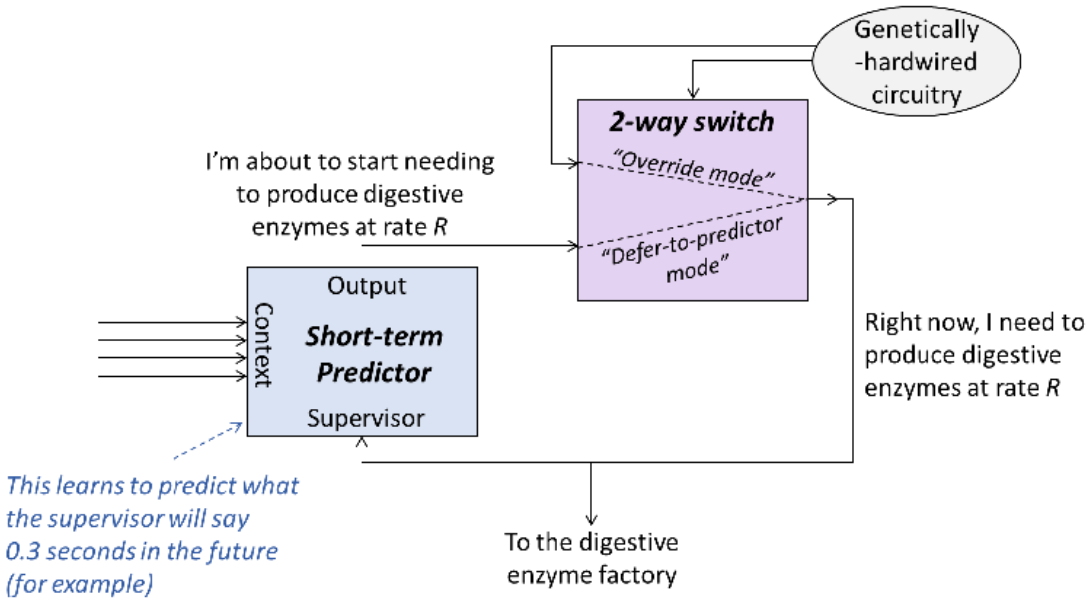
Table of contents:

- Section 5.2 starts with a toy model of a “long-term predictor” circuit, consisting of the “short-term predictor” of the previous post, plus some extra components, wrapped into a closed loop. Getting a good intuitive understanding of this model will be important going forward, and I will walk through how that model would behave under different circumstances.
- Section 5.3 relates that model to Temporal Difference (TD) learning, which is closely related to a “long-term predictor”. I’ll show two variants of the long-term predictor circuit, a “summation” version (which leads to a value function that approximates the sum of future rewards), and a “switch” version (which leads to a value function that approximates the *next* reward, whenever it should arrive, which may not be for a long time). The “summation” version is universal in AI literature, but I’ll suggest that the “switch” version is probably closer to what happens in the brain. Incidentally, these two models are equivalent in cases like AlphaGo, wherein reward arrives in a lump sum right at the end of each episode (= game of Go).
- Section 5.4 will relate long-term predictors to the neuroanatomy of (part of) the telencephalon and brainstem.
    - For the “vertical” neuroanatomy,[1] I’ll describe how the brain houses a huge number of parallel “cortico-basal ganglia-thalamo-cortical loops”, and I’ll suggest that some of these loops function as short-term predictors, with a dopamine signal as supervisor.
    - For the “horizontal” neuroanatomy, I’ll propose that the supervised learning I’m talking about involves (for example) the medial prefrontal cortex, ventral striatum, anterior insular cortex, and amygdala.

- Section 5.5 will offer six lines of evidence that lead me to believe this story: (1) It's a sensible way to implement a biologically-useful capability; (2) It's introspectively plausible; (3) It's evolutionarily plausible; (4) It offers a reconciliation between the "visceromotor" and "motivational" ways to describe the medial prefrontal cortex; (5) It explains the Dead Sea Salt experiment; and (6) It offers a nice explanation of the diversity of dopamine neuron activity.

5.2 Toy model of a "long-term predictor" circuit

A "long-term predictor" is ultimately nothing more than a short-term predictor whose output signal helps determine its own supervisory signal. Here's a toy model of what that can look like:



Toy model of a long-term prediction circuit. I'll spend the next couple subsections walking through how this works. ***Edited to add:*** For this and all similar diagrams in this post, every block at every moment is running in parallel, and likewise every arrow at every moment is carrying a numerical value. So this is *NOT* a control-flow diagram for serial code; rather, it's the kind of diagram you might see describing an FPGA, for example.

- The blue box is the short-term predictor of the previous post. It optimizes its output signal such that it approximates what the supervisor signal will be in 0.3 seconds (as an example).
- The purple box is a 2-way switch. The toggle on the switch is controlled by genetically-hardwired circuitry (gray oval), according to the following rules:
  1. By and large, the switch is in the bottom setting ("defer-to-predictor mode"). This setting is akin to the genetically-hardwired circuitry "trusting" that the short-term predictor's output is sensible, and in particular producing the suggested amount of digestive enzymes.
  2. If the genetically-hardwired circuitry gets a signal that I'm eating something *right now*, and that I don't have adequate digestive enzymes, it flips the switch to "override mode". Regardless of what the short-term predictor says, it sends the signal to manufacture digestive enzymes.
  3. If the genetically-hardwired circuitry has been asking for digestive enzyme production for an extended period, and there's still no food being eaten, then it again

flips the switch to “override mode”. Regardless of what the short-term predictor says, it sends the signal to *stop* manufacturing digestive enzymes.

Note: You can assume that all the signals in the diagram can vary continuously across a range of values (as opposed to being discrete on/off signals), with the exception of the signal that toggles the 2-way switch.[2] In the brain, smoothly-adjustable signals might be created by, for example, rate-coding—i.e., encoding information as the frequency with which a neuron is firing.

5.2.1 Toy model walkthrough part 1: static context

Let's walk through what would happen in this toy model.[3] To start with, ***assume that the "context" is static for some extended period of time***. For example, imagine a situation where some ancient worm-like creature is digging in the sandy ocean bed for many consecutive minutes. Plausibly, its sensory environment would stay pretty much constant as long as it keeps digging, as would its thoughts and plans (insofar as this ancient worm-like creature has "thoughts and plans" in the first place). Or if you want another example of (approximately) static context—this one involving a human rather than a worm—hang on until the next subsection.

In the static-context case, let's first consider what happens when the switch is sitting in "defer-to-predictor mode": Since the output is looping right back to the supervisor, there is no error in the supervised learning module. The predictions are correct. The synapses aren't changing. Even if this situation is very common, it has no bearing on how the short-term predictor eventually winds up behaving.

The times that *do* matter for the eventual behavior of the short-term predictor are those rare times that we go into "override mode". Think of the overrides as like a sporadic "injection of ground truth". They produce an error signal in the short-term predictor's learning algorithm, changing its adjustable parameters (e.g. synapse strengths).

After enough life experience (a.k.a. "training" in ML terminology), the short-term predictor should have the property that *the overrides balance out*. There may still be occasional overrides that *increase* digestive-enzyme production, and there may still be occasional overrides that *decrease* digestive-enzyme production, but those two types of overrides should happen with similar frequency. After all, if they *didn't* balance out, the short-term predictor's internal learning algorithm would gradually change its parameters so that they *did* balance out.

And that's just what we want! We'll wind up with appropriate digestive enzyme production at appropriate times, in a way that properly accounts for any information available in the context data—what the animal is doing right now, what it's planning to do in the future, what its current sensory inputs are, etc.

5.2.1.1 David-Burns-style exposure therapy—a possible real-life example of the toy model with static context?

As it happens, I recently read David Burns's book *Feeling Great* (my review). David Burns has a very interesting approach to exposure therapy—an approach that happens to serve as an excellent example of how my toy model works in the static-context situation!

Here's the short version. (Warning: If you're thinking of doing exposure therapy on yourself at home, *at least* read the whole book first!) Excerpt from the book:

> For example, when I was in high school, I wanted to be on the stage crew of *Brigadoon*, a play my school was putting on, but it required overcoming my fear of heights since the stage crew had to climb ladders and work near the ceiling to adjust the lights and curtains. My drama teacher, Mr. Krishak, helped me overcome this fear with the very type of exposure techniques I'm talking about. He led me to the theater and put a tall ladder in the middle of the stage, where there was nothing nearby to grab or hold on to. He told me all I had to do was stand on the top of the ladder until my fear disappeared. He reassured me that he'd stand on the floor next to me and wait.
>
> I began climbing the ladder, step by step, and became more and more frightened. When I got to the top, I was terrified. My eyes were almost 18 feet from the floor, since the ladder was 12 feet tall, and I was just over 6 feet tall. I told Mr. Krishak I was in a panic and asked what I should do. Was there something I should say, do, or think about to make my anxiety go away? He shook his head and told me to just stand there until I was cured.

I continued to stand there in terror for about ten more minutes. When I told Mr. Krishak I was still in a panic, he assured me that I was doing great and that I should just stand there a few more minutes until my anxiety went away. A few minutes later, my anxiety suddenly disappeared. I couldn't believe it!

I told him, "Hey, Mr. Krishak, I'm cured now!"

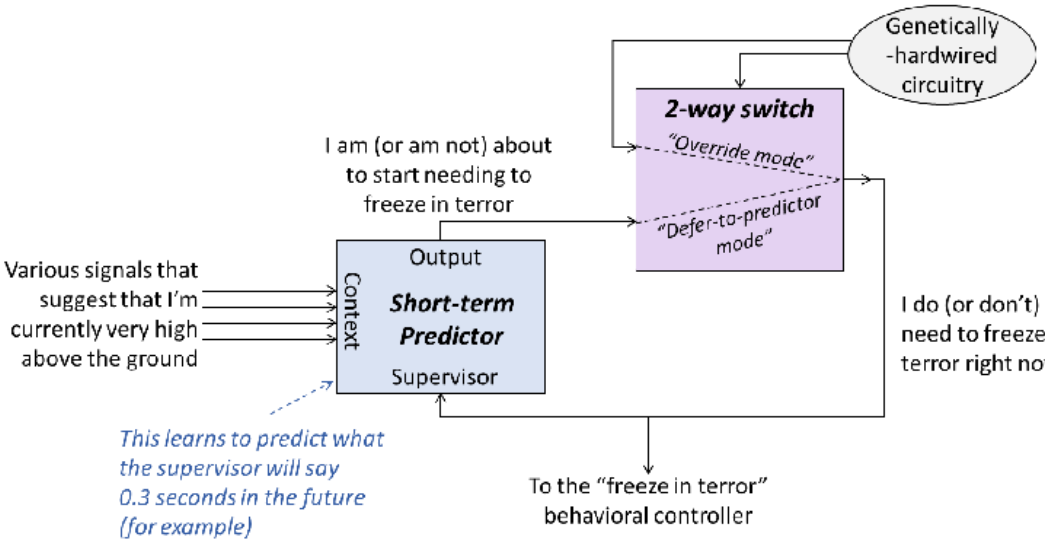
He said, "Great, you can come on down from the ladder now, and you can be on the stage crew of *Brigadoon*!"

I had a blast working on the stage crew. I absolutely loved climbing ladders and adjusting the lights and curtains near the ceiling, and I couldn't even remember why or how I'd been so afraid of heights.

This story seems to be beautifully consistent with my toy model here. David started the day in a state where his short-term-predictors output "extremely strong fear reactions" when he was up high. As long as David stayed up on the ladder, those fear-reaction short-term-predictors kept on getting the same context data, and therefore they kept on firing their outputs at full strength. And David just kept feeling terrified.

Then, after 15 boring-yet-terrifying minutes on the ladder, some innate circuit in David's brainstem issued an *override*—as if to say, "C'mon, nothing is changing, nothing is happening, we can't just keep burning all these calories all day. It's time to calm down now." The short-term-predictors continued sending the same outputs as before, but the brainstem exercised its veto power, and forcibly reset David's cortisol, heart-rate, etc., back to baseline. This "override" state immediately created *error signals* in the relevant short-term-predictors in David's amygdala! And the error signals, in turn, led to model updates! The short-term predictors were all edited, and from then on, David was no longer afraid of heights.

This story kinda feels like speculation piled on top of speculation, but whatever, I happen to think it's right. If nothing else, it's good pedagogy! Here's the diagram for this situation; make sure you can follow all the steps.



5.2.2 Toy model walkthrough, assuming changing context

The previous subsections assumed static context lines (constant sensory environment, constant behaviors, constant thoughts and plans, etc.). What happens if the context is *not* static?

If the context lines are changing, then it's no longer true that learning happens *only* at "overrides". If context changes in the absence of "overrides", it will result in changing of the output, and *the new output will be treated as ground truth for what the old output should have been*. Again, this seems to be just what we want: if we learned something new and relevant in the last second, then our current expectation should be more accurate than our previous expectation, and thus we have a sound basis for updating our models.

5.3 Value function calculation (TD learning) as a special case of long-term prediction

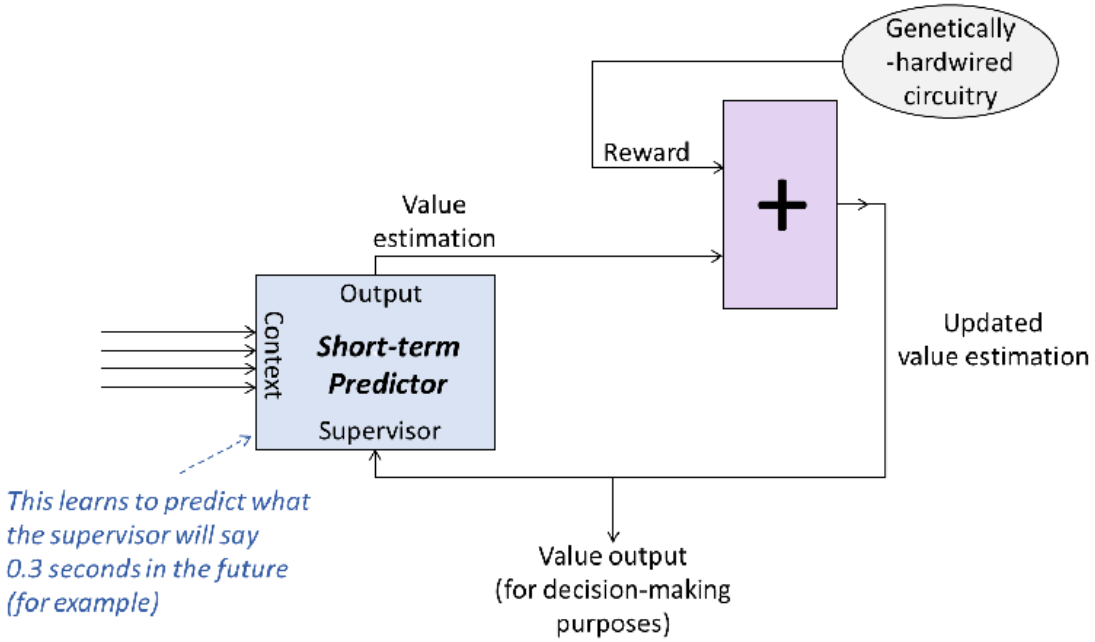
At this point, ML experts will recognize a resemblance to Temporal Difference (TD) learning. It's not quite the same, though. The differences are:

First, TD learning is usually used in reinforcement learning (RL) as a method for going from a reward function to a value function. By contrast, I was talking about things like "digestive enzyme production", which are neither rewards nor values.

In other words, there is a generally-useful motif that involves going from some immediate quantity X to "long term expectation of X". The calculation of a value function from a reward function is an *example* of that motif, but it's not the only useful example.

(As a matter of terminology, it seems to be generally accepted that the term "TD learning" can in fact apply to things that are *not* RL value functions.[4] However, empirically in my own experience, as soon as I mention "TD learning", the people I'm talking to immediately assume I *must* be talking about RL value functions. So I want to be clear here.)

Second, to get something closer to traditional TD learning, we'd need to replace the 2-way *switch* with a 2-way *summation*—and then the "overrides" would be analogous to rewards. Much more on "switch vs summation" in the next subsection.



Here's a TD learning circuit that would behave similarly to what you'd see in an AI textbook. Note the purple box on the right: compared to the previous figure, I replaced the 2-way

switch with a 2-way *summation*. More on “switch vs summation” in the next subsection.

Third, there are many additional ways to tweak the circuit which are frequently used in AI textbooks, and some of those may be involved in the brain circuits too. For example, we can put in time-discounting, or different emphases on false-positives vs false-negatives (see my discussion of distributional learning in Section 5.5.6.1 below), etc.

To keep things simple, I will be ignoring all these possibilities (including time-discounting) in the discussion below.

5.3.1 Switch (i.e., value = expected next reward) vs summation (i.e., value = expected sum of future rewards)?

The figures above show two variants of our toy model. In one, the purple box is a two-way *switch* between “defer to the short-term predictor” and some independent “ground truth”. In the other, the purple box is a two-way *summation* instead.

The *switch* version trains the short-term-predictor to predict the next ground truth, whenever it should arrive.

The *summation* version trains the short-term-predictor to predict the *sum* of future ground truth signals.

The correct answer could also be “something in between switch and summation”. Or it could even be “none of the above”.

RL papers universally use the summation version—i.e., “value is the expected sum of future rewards”. What about biology? And which is actually better?

It doesn’t always matter! Consider AlphaGo. Like every RL paper today, AlphaGo was originally formulated in the summation paradigm. But it happens to have one and only one nonzero reward signal per game, namely +1 at the end of the game if it wins, or -1 if it loses. In that case, switch vs summation makes no difference. The only difference is one of terminology:

- In the summation case, we would say “each non-terminal move in the Go game has reward=0”.
- In the switch case, we would say “each non-terminal move in the Go game has a reward of (null)”.

(Do you see why?)

But in other cases, it does matter. So back to the question: should it be switch or summation?

Let’s step back. What are we trying to do here?

One thing that a brain needs to do is make decisions that weigh cross-domain tradeoffs. If you’re a human, you need to decide whether to watch TV or go to the gym. If you’re some ancient worm-like creature, you need to “decide” whether to dig or to swim. Either way, this “decision” impacts energy balance, salt balance, probability of injury, probability of mating—you name it. The design goal in the decision-making algorithm is that you make the decision that maximizes inclusive genetic fitness. How might that goal be best realized?

One method involves building a *value function* that estimates the organism’s inclusive genetic fitness (compared to some arbitrary—indeed, possibly time-varying—baseline), conditional on continuing to execute a given course of action. Of course it won’t be a perfect estimate—*real* inclusive genetic fitness can only be calculated in hindsight, many generations after the fact. But once we have such a value function, however imperfect, we can plug it into an algorithm that

makes decisions to maximize value (more on this in the next post), and thus we get approximately-fitness-maximizing behavior.

So having a value function is key for making good decisions that weigh cross-domain tradeoffs. But nowhere in this story is the claim "value is the expectation of a sum of future rewards"! That's a particular way of setting up the value-approximating algorithm, a method which might or might not be well suited to the situation at hand.

I happen to think that brains use something closer to the switch circuit, not the summation circuit, not only for homeostatic-type predictions (like the digestive enzymes example above), but also for value functions, contrary to mainstream RL papers. Again, I think it's really "neither of the above" in all cases; just that it's *closer* to switch.

Why do I favor "switch" over "summation"?

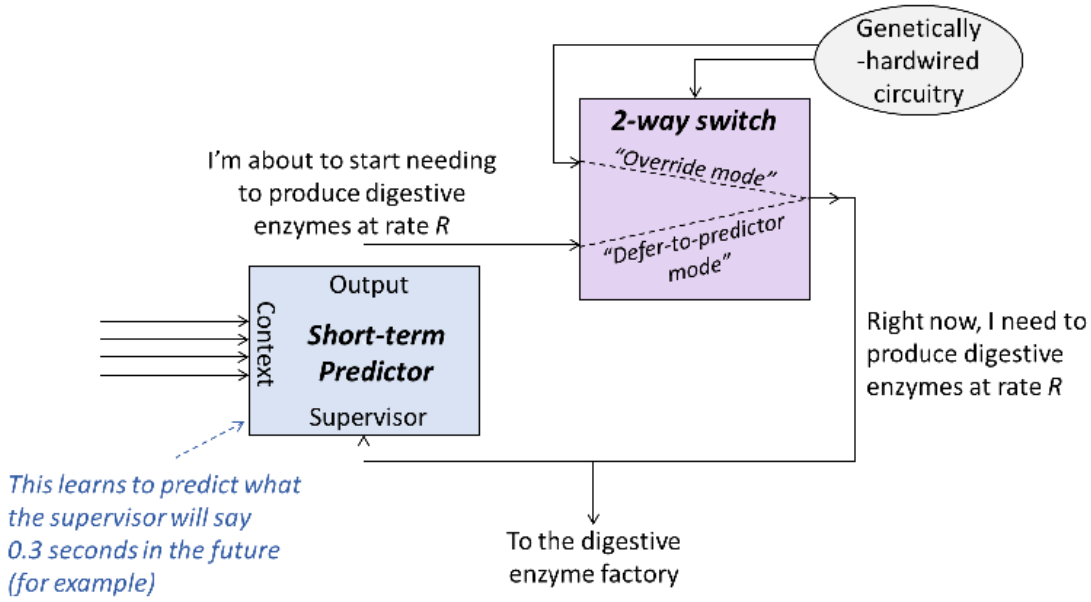
An example: sometimes I stub my toe and it hurts for 20 seconds; other times I stub my toe and it hurts for 40 seconds. But I don't think of the latter as *twice as bad* as the former. In fact, even five minutes later, I wouldn't remember which is which. (See the peak-end rule.) This is the kind of thing I would naturally expect from switch, but is an awkward fit for summation. It's not *strictly* incompatible with summation; it just requires a more complicated, value-dependent reward function. As a matter of fact, if we allow the reward function to depend on value, then switch and summation can imitate each other.

Anyway, in upcoming posts, I'll be assuming switch, not summation. I don't think it matters very much for the big picture. I *definitely* don't think it's part of the "secret sauce" of animal intelligence, or anything like that. But it does affect some of the detailed descriptions.

The next post will include more details of reinforcement learning in the brain, including how "reward prediction error" works and so on. I am bracing for lots of confused readers, who will be disoriented by the fact that I'm assuming a different relationship between value and reward than what everyone is used to. For example, in my picture, "reward" is a *synonym* for "ground truth for what the value function should be right now"—*both* should account for not only the organism's current circumstances but also its future prospects. Sorry in advance for any confusion! I will do my best to be clear.

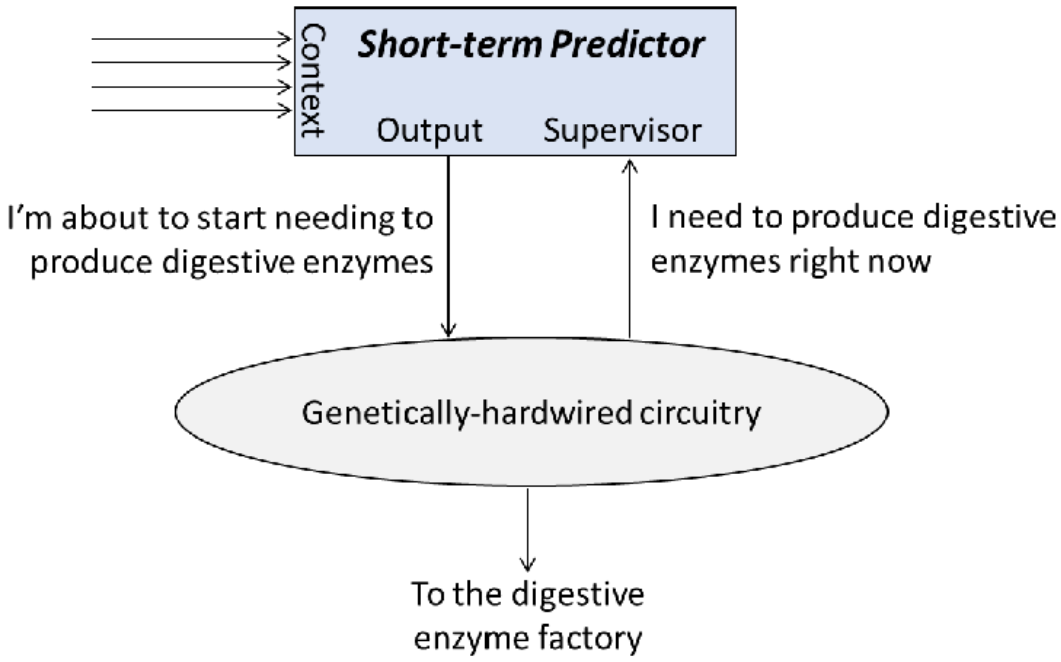
5.4 An array of long-term predictors involving the telencephalon & brainstem

Here's the long-term-predictor circuit from above:



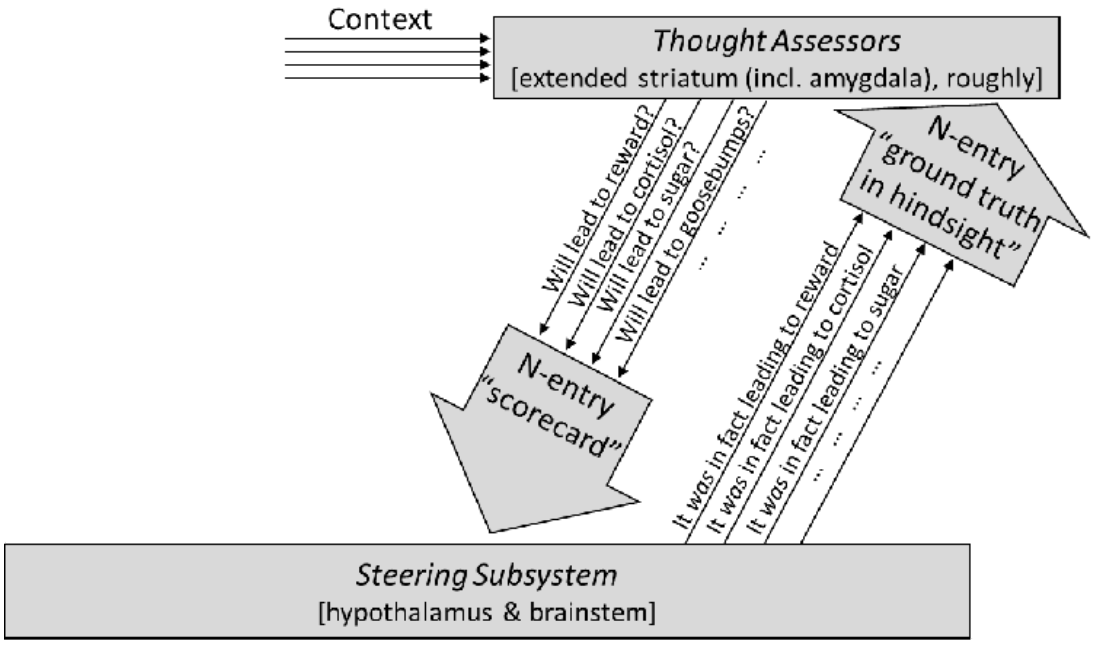
Copied from above.

I can lump together the 2-way switch with the rest of the genetically-hardwired circuitry, and then rearrange the boxes a bit, and I get the following:



Same as above, but drawn differently.

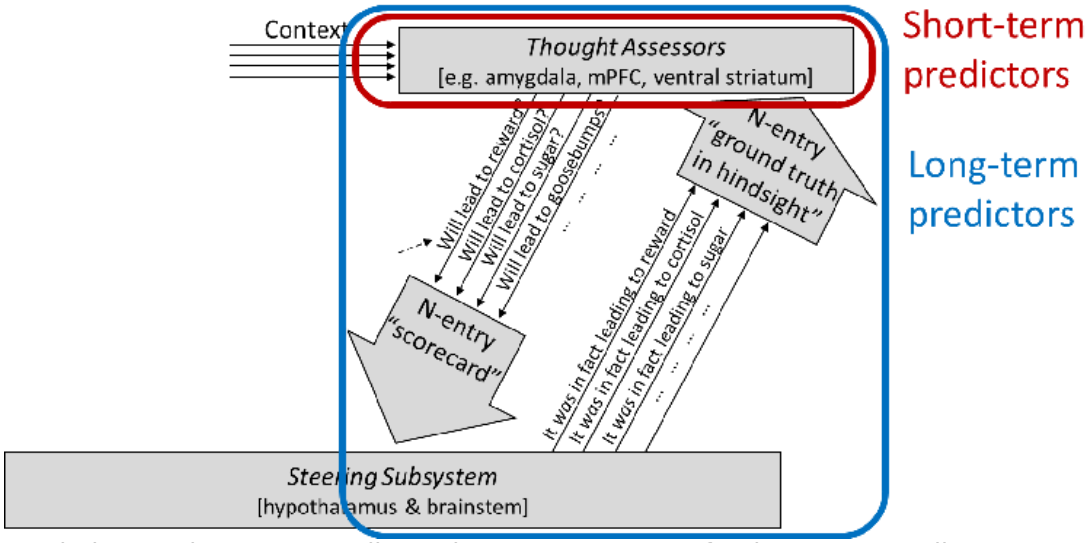
Now, obviously digestive enzymes are just one example. Let's draw in some more examples, add some hypothesized neuroanatomy, and include other terminology. Here's the result:



I claim that there is a bank of long-term-predictors, consisting of an array of short-term-predictors in the telencephalon, each with a closed-loop connection to a corresponding Steering Subsystem circuit. I'm calling the former (telencephalon) part by the name "Thought Assessors", for reasons explained in Section 5.5.4 below.

Excellent! We're halfway to my big picture of decision-making and motivation. The rest of the picture—including the "actor" part of actor-critic reinforcement learning—will come in the next post, and will fill in the hole in the top-left side of that diagram. (The term "Steering Subsystem" comes from Post #3.)

Here's one more diagram and caption for pedagogical purposes.



Reminder: a "short-term predictor" is *one component of* a "long-term predictor". Here's where both those things fit into that diagram above. The *only* thing that makes it a long-term predictor is the possibility of "defer-to-predictor mode"—i.e., the Steering Subsystem might send a "ground truth in hindsight" signal that is not *really* "ground truth" in the normal sense, but is rather a copy of the corresponding entry on the scorecard. In other words, "defer-to-predictor mode" is like the Steering

> Subsystem saying to the short-term predictor: “OK sure, whatever, I’ll take your word for it”. If the Steering Subsystem regularly keeps a signal in “defer-to-predictor mode” for 10 minutes straight, then we can get predictions that anticipate the future by up to 10 minutes. Conversely, if the Steering Subsystem *never* uses “defer-to-predictor mode” for a certain signal, then we shouldn’t really be calling it a “long-term predictor” in the first place.

In the next two subsections, I will elaborate on the neuroanatomy which I’m hinting at in this diagram, and then I’ll talk about why you should believe me.

5.4.1 “Vertical” neuroanatomy:[1] cortico-basal ganglia-thalamo-cortical loops

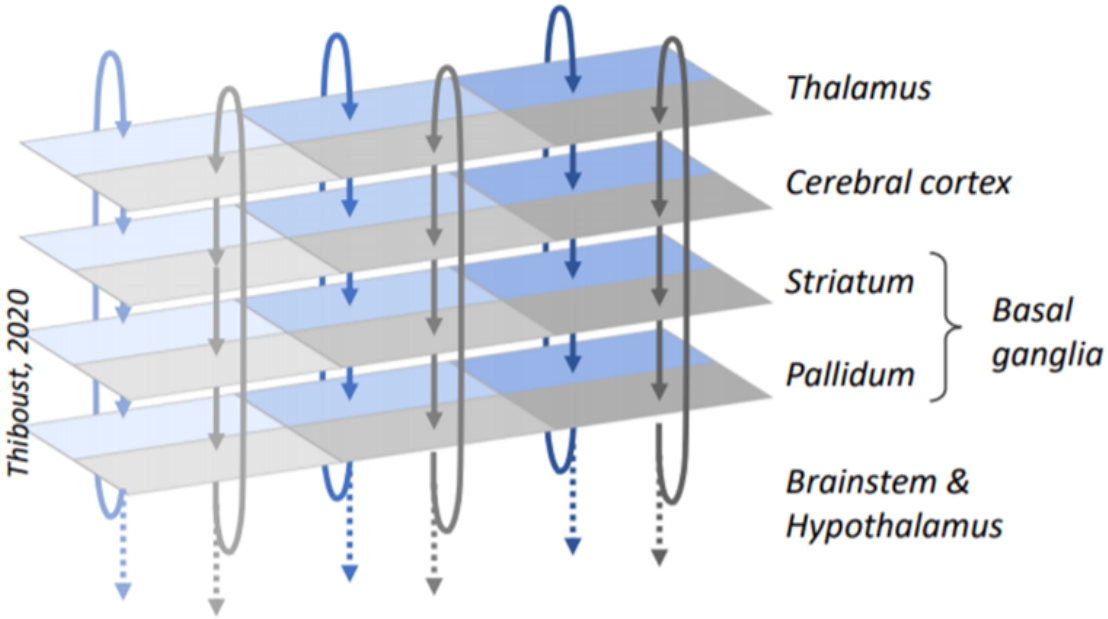
In my post Big Picture of Phasic Dopamine, I talked about the theory (due originally to Larry Swanson) that the *whole telencephalon* is nicely organized into three layers (cortex, striatum, pallidum):

Cortex-like part of the loops	Hippo-campus	Amygdala [basolateral part]	Piriform cortex	Medial prefrontal cortex	Motor & “planning” cortex
Striatum-like part of the loops	Lateral septum	Amygdala [central part]	Olfactory tubercle	Ventral striatum	Dorsal striatum
Pallidum-like part of the loops	Medial septum	BNST	Substantia innominata	Ventral pallidum	Globus pallidus

The entire telencephalon—neocortex, hippocampus, amygdala, everything—can be divided into cortex-like structures, striatum-like structures, and pallidum-like structures. If two structures are in the same column in this table, that means they’re wired together into cortico-basal ganglia-thalamo-cortical loops (see next paragraph). This table is incomplete and oversimplified; for a better version see Fig. 4 here .

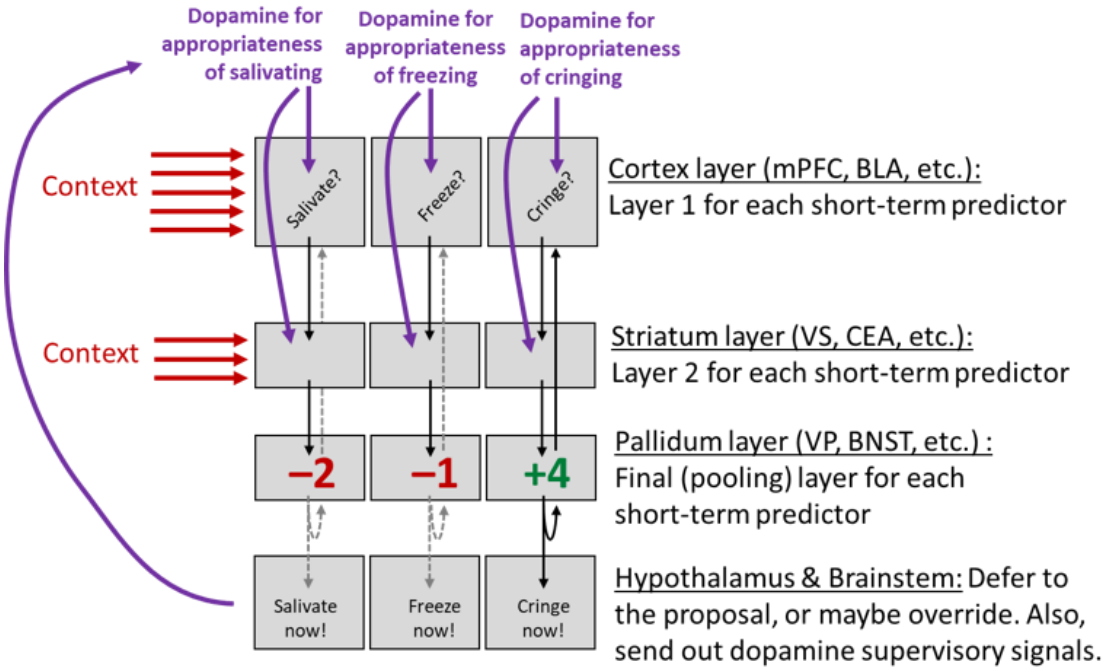
This idea then connects to the earlier (and now widely accepted) theory, dating to Alexander 1986, that these three layers of the telencephalon are interconnected into a large number of parallel “cortico-basal ganglia-thalamo-cortical loops”, which can be found in almost every part of the telencephalon.

Here’s a little illustration:



Simplified cartoon illustration of how the brain has many parallel cortico-basal ganglia-thalamo-cortical loops. Source: Matthieu Thiboust.

Given all that, here is a possible rough model for how this loop architecture relates to the short-term predictor learning algorithm that I've been talking about:



WARNING: DON'T TAKE THIS DIAGRAM TOO LITERALLY. See Big Picture of Phasic Dopamine for *slightly* more details, but mostly I haven't looked into it much, and in particular the "Layer 1, Layer 2, Final (pooling) layer" labels are kinda just spitballing. (The "pooling" is based on there being 2000× more neurons in the striatum than the pallidum—see here.) Acronyms: BLA=basolateral amygdala, BNST=bed nucleus of the stria terminalis, CEA=central amygdala, mPFC=medial prefrontal cortex, VP=ventral pallidum, VS=ventral striatum.

5.4.2 “Horizontal” neuroanatomy—cortical specialization

The previous subsection was about the “vertical” three-layer structure of the telencephalon. Now let’s switch to the “horizontal” structure, i.e. the fact that different parts of the cortex do different things (in cooperation with the corresponding parts of the striatum and pallidum).

This is oversimplified, but here’s my latest attempt at (part of) the cortex in a nutshell:

- The extended motor cortex (and corresponding striatum) is the cortex’s main output region for behaviors involving skeletal muscles, like reaching and walking.
- The medial prefrontal cortex (mPFC—which also includes anterior cingulate cortex) (and corresponding (ventral) striatum) is the cortex’s main output region for behaviors involving autonomic / visceromotor / hormonal actions, like releasing cortisol, vasoconstriction, goosebumps, and so on.
- The amygdala (which has both cortex-like and striatum-like parts) is the cortex’s main output region for certain behaviors that involve *both* skeletal muscle actions *and* autonomic actions, like flinching-reactions, or freezing-reactions (when frightened), and so on.
- The insular cortex (and corresponding (ventral) striatum) is the cortex’s main *input* region for autonomic / homeostatic / body status information, like blood sugar levels, pain, cold, taste, muscle strain, etc.

I won’t talk about the motor cortex in this series, but I think the other three are all involved in these long-term prediction circuits. For example:

- I claim that if you look at a little subregion in the medial prefrontal cortex, you might find that it’s being trained to fire in proportion to the probability of upcoming cortisol release;
- I claim that if you look at a little subregion in the amygdala, you might find that it’s being trained to fire in proportion to the probability of upcoming freezing-reactions;
- I claim that if you look at a little subregion of the (anterior) insular cortex, you might find that it’s being trained to fire in proportion to the probability of upcoming cold feelings in your left arm.

5.5 Six reasons I like this “array of long-term predictors” picture

5.5.1 It’s a sensible way to implement a biologically-useful capability

If you start producing digestive enzymes *before* eating, you’ll digest faster. If your heart starts racing *before* you see the lion, then your muscles will be primed and ready to go when you *do* see the lion. Etc.

So these kinds of predictors seem obviously useful.

Moreover, as discussed in the previous post (Section 4.5.2), the technique I’m proposing here (based on supervised learning) seems either superior to or complementary with other ways to meet these needs.

5.5.2 It’s introspectively plausible

For one thing, we *do* in fact start salivating before we eat the cracker, start feeling nervous before we see the lion, etc.

For another thing, consider the fact that all the actions I'm talking about in this post are *involuntary*: you cannot salivate on command, or dilate your pupils on command, etc., at least not in quite the same way that you can wiggle your thumb on command.

(More on voluntary actions in the next post—they're in a whole different part of the telencephalon.)

I'm glossing over a bunch of complications here, but the involuntary nature of these things seems pleasingly consistent with the idea that they are being trained by their own dedicated supervisory signals, straight from the brainstem. They're slaves to a different master, so to speak. We can kinda *trick them* into behaving in certain ways, but our control is limited and indirect.

5.5.3 It's evolutionary plausible

As discussed in Section 4.4 of the previous post, the simplest short-term predictor is *extraordinarily* simple, and the simplest long-term predictor is only a bit more complicated than that. And these very simple versions are already plausibly fitness-enhancing, even in very simple animals.

Moreover, as I discussed a while back (Dopamine-supervised learning in mammals & fruit flies), there is an array of little learning modules in the fruit fly, playing a seemingly-similar role to what I'm talking about here. Those modules *also* use dopamine as a supervisory signal, and there is some genomic evidence of a homology between those circuits and the mammalian telencephalon.

5.5.4 It offers a reconciliation between "visceromotor" and "motivation" pictures of the medial prefrontal cortex (mPFC)

Take the mPFC (which also includes the anterior cingulate cortex—ACC), as an example. People talk about this region in two quite different ways:

- On the one hand, as mentioned above (Section 5.4.2), mPFC is described as a visceromotor / homeostatic / autonomic motor output region—it issues commands to control hormones, to execute sympathetic and parasympathetic nervous system reactions, and so on. For example, "electrical stimulation of the infralimbic cortex has been shown to affect gastric motility and to cause hypotension", or this paper says stimulating mPFC caused "[bristling]; pupillary dilation; and changes in blood pressure, respiratory rate, and heart rate", or see Bud Craig's book which characterizes ACC as a homeostatic motor output center. This way of thinking also elegantly explains the fact that the region is agranular (missing layer #4 out of the 6 neocortex layers), which implies "output region" both for theoretical reasons and by analogy with the (agranular) motor cortex.
- On the other hand, mPFC is frequently described as being related to a host of vaguely-motivation-related activities. For example, Wikipedia mentions "attention allocation, reward anticipation, decision-making, ethics and morality, impulse control … and emotion" in regards to ACC.

I think my picture works for both:[5]

For the first (visceromotor) perspective, if you look at Section 5.2 above, you'll see that the predictors' outputs *do* in fact cause homeostatic changes—at least, they do when the genetically-hardwired circuitry of the Steering Subsystem has set that signal in "defer-to-predictor mode" (as opposed to "override mode").

For the second (motivation) perspective, this will make a bit more sense after the next post, but note my suggestive description of a "scorecard" in the diagram of Section 5.4. The idea is: The

“context” lines going into the “Thought Assessors” contain the horrific complexity of *everything in your conscious mind and more*—where you are, what you’re seeing and doing, what you’re thinking about, what you’re planning to do in the future and why, etc. The relatively simple, genetically-hardcoded Steering Subsystem can’t make heads or tails of any of that!

But that’s a dilemma, because the Steering Subsystem is the source of rewards / drives / motivations! How can the Steering Subsystem issue rewards for a good plan, if it can’t make heads or tails of what you’re planning??

The “scorecard” is the answer. It takes all that horrific complexity and distills it into a nice standardized scorecard—exactly the kind of thing that genetically-hardcoded circuits in the Steering Subsystem can easily process.

Thus, whenever there’s an interaction between thoughts and drives—emotions, decision-making, ethics, aversions, etc.—the “Thought Assessors” need to be involved as an intermediary.

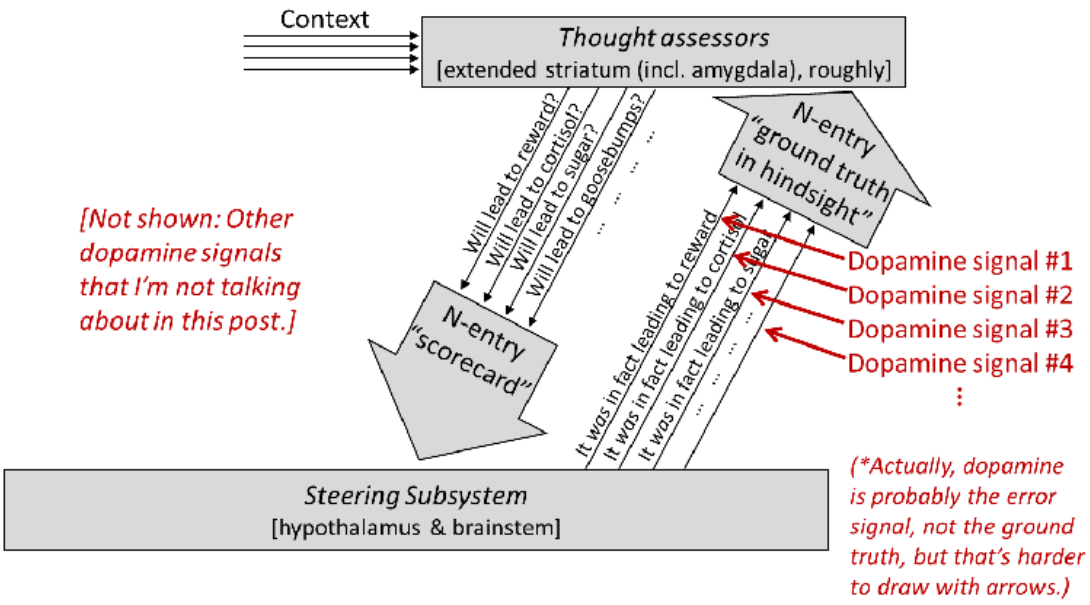
5.5.5 It explains the Dead Sea Salt Experiment

See my discussion in my old post Inner alignment in salt-starved rats. In brief, experimenters sporadically played a sound and popped an object into a rat’s cage, and immediately thereafter sprayed super-salty water directly into the rat’s mouth. The rat found the saltwater disgusting, and started reacting with horror to the sound and object. Then later, the experimenters made the rat feel salt-deprived. When they played the sound and popped the object *this* time, the rat got very excited—even though the rat had never been salt-deprived before in its life.

In our setup, this is exactly what we expect: when the sound and object appear, the “I anticipate tasting salt” predictor starts firing like crazy. Meanwhile the Steering Subsystem (hypothalamus & brainstem) has hardwired circuitry that says “If I’m salt-deprived, and if the ‘scorecard’ from the Learning Subsystem suggests that I will soon taste salt, then *that’s awesome*, and whatever thought the Learning Subsystem is thinking, it should pursue that idea with gusto!”

5.5.6 It offers a nice explanation for (some of) the diversity of dopamine neuron activity

Recall from Section 5.4.1 above that I’m claiming that dopamine neurons carry the supervisory signals of all these supervised-learning modules.[6]



There's a pop-science misconception that there is a (singular) dopamine signal in the brain, and it bursts when good things are happening. In reality, there are many different dopamine neurons doing many different things.

Thus we get the question: what are all these diverse dopamine signals doing? There's no consensus; claims in the literature are all over the place. But I can throw my hat into the ring: in my picture described above, there are probably hundreds or thousands of short-term predictors in the telencephalon, predicting hundreds or thousands of different things, and *they each need a different dopamine supervisory signal*!

(And there are even *more* dopamine signals besides those! One such signal, associated with the brain's "main" reward prediction error signal, will show up in the next post. Still others are off-topic for this series but discussed here.)

If my story is right, what would we expect to see in dopamine-measuring experiments?

Imagine a rat running through a maze. Moment by moment, its array of predictors are getting dopamine supervisory signals about its various hormone levels, its heart rate, its expectation of drinking and eating and having a sore leg and freezing and tasting salt, and on and on. In short, we expect dopamine neurons to be bouncing up and down in all kinds of different ways.

Thus, pretty much any instance where an experimenter has measured that a dopamine neuron is correlated with some behavioral variable, it's *probably* consistent with my picture too.

Here are a couple examples:

- There are dopamine neurons that burst for salient stimuli like unexpected flashes of light (ref). Can I explain that? Sure, no problem! I say: they could be supervisory signals saying "this would have been a good time to orient", or "to flinch", or "to raise your heart rate", etc.
- There are dopamine neurons that correlate with the velocity with which a mouse is running on a treadmill-ball (ref). Can I explain that? Sure, no problem! I say: they could be supervisory signals saying "expect sore muscles", or "expect cortisol", or "expect high heart rate", etc.

Here's another data point which seems reassuringly consistent with my picture. A few dopamine neurons burst when *aversive* things happen (ref). Four of the five regions[7] in which such neurons can be found (according to the linked paper) are right where I expect that array of short-

term predictors to be—namely, the cortex-like and striatum-like layers of amygdala, and medial prefrontal cortex (mPFC), and the ventromedial shell of the nucleus accumbens, which is (at least roughly?) the striatum stop of the mPFC cortico-basal ganglia-thalamo-cortical loops. This is exactly what I expect in my picture. For example, if a mouse gets shocked, then a "should-I-freeze-now" predictor would get a supervisory signal saying "Yes, you should have been freezing".

Side note: Lammel et al. 2014 mentions so-called "'non-conventional' VTA [dopamine] neurons" in "medial posterior VTA (PN and medial PBP)". These seem to project to exactly the non-value-function Thought Assessor areas, and it's claimed that they have different firing patterns from other dopamine neurons. Maybe the firing pattern difference is reflective of the different requirements of supervised learning versus reinforcement learning? (I'm not an expert; I'm just flagging that it sounds intriguing and would be worth looking into more.)

UPDATE JAN 2023: Upon further investigation (thanks Nathaniel Daw), I think what I'm talking about here is basically the right explanation for the diverse dopamine signals on the fringes of VTA / SNc, or something like that, but the fine-grained dopamine diversity more typically measured has a different explanation which is at least spiritually closer to the "distributional" story next.

5.5.6.1 Aside: Distributional predictor outputs

I didn't talk about it in the last post, but short-term predictors have hyperparameters in their learning algorithms, two of which are "how strongly to update upon a false-positive (overshoot) error", and "how strongly to update upon a false-negative (undershoot) error". As the ratio of these two hyperparameters varies from 0 to ∞, the resulting predictor behavior varies from "fire the output if there's even the faintest chance that the supervisor will fire" to "never fire the output unless it's all but certain that the supervisor will fire".

Therefore, if we have *many* predictors, each with a different ratio of those hyperparameters, then we can (at least approximately) output a *probability distribution* for the prediction, rather than a point estimate.

A recent set of experiments from DeepMind and collaborators found evidence (based on measurements of dopamine neurons) that the brain does in fact use this trick, at least for reward prediction.

I speculate that it may use the same trick for the other long-term predictors too—e.g. maybe the predictions of arm pain and cortisol and goosebumps etc. are *all* in the form of ensembles of long-term predictors that each sample a probability distribution.

I bring this up, first, because it's another example where dopamine neurons are behaving in a way that seems pleasingly consistent with my worldview, and second, because it's plausibly useful for AGI safety—and thus I was looking for an excuse to bring it up anyway!

5.6 Conclusion

Anyway, as usual I don't pretend to have smoking-gun proof of my hypothesis (i.e. that the brain has an array of long-term predictors involving telencephalon-brainstem loops), and there are some bits that I know I'm still confused about. But considering the evidence in the previous subsection (and rest of the post), I wind up feeling strongly that I'm broadly on the right track. I'm happy to discuss more in the comments. Otherwise, onward to the next post, where we will finally put everything together into a big picture of how I think motivation and decision-making work in the brain!

1. ^

   'Horizontal' neuroanatomy versus 'vertical' neuroanatomy is my idiosyncratic terminology, but I'm hoping it's intuitive. If you imagine stretching out the cortex into a sheet, oriented

horizontally, then the ‘vertical’ neuroanatomy would include e.g. the interconnections between cortical and subcortical structures, and the ‘horizontal’ neuroanatomy would include e.g. the different roles played by different parts of the cortex. See also the table in Section 5.4.1.

2. ^

   To be clear, in reality, there probably isn’t a discrete all-or-nothing 2-way switch here. There could be a “weighted average” setting, for example. Remember, this whole discussion is just a pedagogical “toy model”; I expect that reality is more complicated in various respects.

3. ^

   I note that I’m just running through this algorithm in my head; I haven’t simulated it. I’m optimistic that I didn’t majorly screw up, i.e. that everything I’m saying about the algorithm is qualitatively true, or at least *can* be qualitatively true with appropriate parameter settings and perhaps other minor tweaks.

4. ^

   Examples of using the terminology “TD learning” for something which is not related to RL reward functions include “TD networks”, and the Successor Representations literature (example), or this paper, etc.

5. ^

   The classic attempt to reconcile “visceromotor” and “motivation” pictures of mPFC is Antonio Damasio’s “somatic marker hypothesis”. My discussion here has some similarities and some differences from the somatic marker hypothesis. I won’t get into that; it’s off-topic.

6. ^

   As in the previous post, when I say that “dopamine carries the supervisory signal”, I’m open to the possibility that dopamine is actually a closely-related signal like the error signal, or the negative error signal, or the negative supervisory signal. It really doesn’t matter for present purposes.

7. ^

   The fifth area where that paper found dopamine neurons bursting under aversive circumstances, namely the tail of the striatum, has a different explanation I think—see here.

[Intro to brain-like-AGI safety] 6. Big picture of motivation, decision-making, and RL

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

6.1 Post summary / Table of contents

Part of the "Intro to brain-like-AGI safety" post series .

Thus far in the series, Post #1 set out some definitions and motivations (what is "brain-like AGI safety" and why should we care?), and Posts #2 & #3 split the brain into a Learning Subsystem (telencephalon and cerebellum) that "learns from scratch" using learning algorithms, and a Steering Subsystem (hypothalamus and brainstem) that is mostly genetically-hardwired and executes innate species-specific instincts and reactions.

Then in Post #4, I talked about the "short-term predictor", a circuit which learns, via supervised learning, to predict a signal in advance of its arrival, but only by perhaps a fraction of a second. Post #5 then argued that if we form a closed loop involving *both* a set of short-term predictors in the Learning Subsystem *and* a corresponding set of hardwired circuits in the Steering Subsystem, we can get a "long-term predictor". I noted that the "long-term predictor" circuit is closely related to temporal difference (TD) learning.

Now in this post, we fill in the last ingredients—roughly the "actor" part of actor-critic reinforcement learning (RL)—to get a whole big picture of motivation and decision-making in the human brain. (I'm saying "human brain" to be specific, but it would be a similar story in any other mammal, and to a lesser extent in any vertebrate.)

The reason I care about motivation and decision-making is that if we eventually build brain-like AGIs (cf. Post #1), we'll want to build them so that they have some motivations (e.g. being helpful) and not others (e.g. escaping human control and self-reproducing around the internet). Much more on that topic in later posts.

Teaser for upcoming posts: The next post (#7) will walk through a concrete example of the model in this post, where we can watch an innate drive lead to the formation of an explicit goal, and adoption and execution of a plan to accomplish it. Then starting in Post #8 we'll switch gears, and from then on you can expect substantially *less* discussion of neuroscience and *more* discussion of AGI safety (with the exception of one more neuroscience post towards the end).

Unless otherwise mentioned, everything in this post is "things that I believe right now", as opposed to neuroscience consensus. *(Pro tip: there is never a neuroscience consensus.)* Relatedly, I will make minimal effort to connect my hypotheses to others in the literature, but I'm happy to chat about that in the comments section or by email.

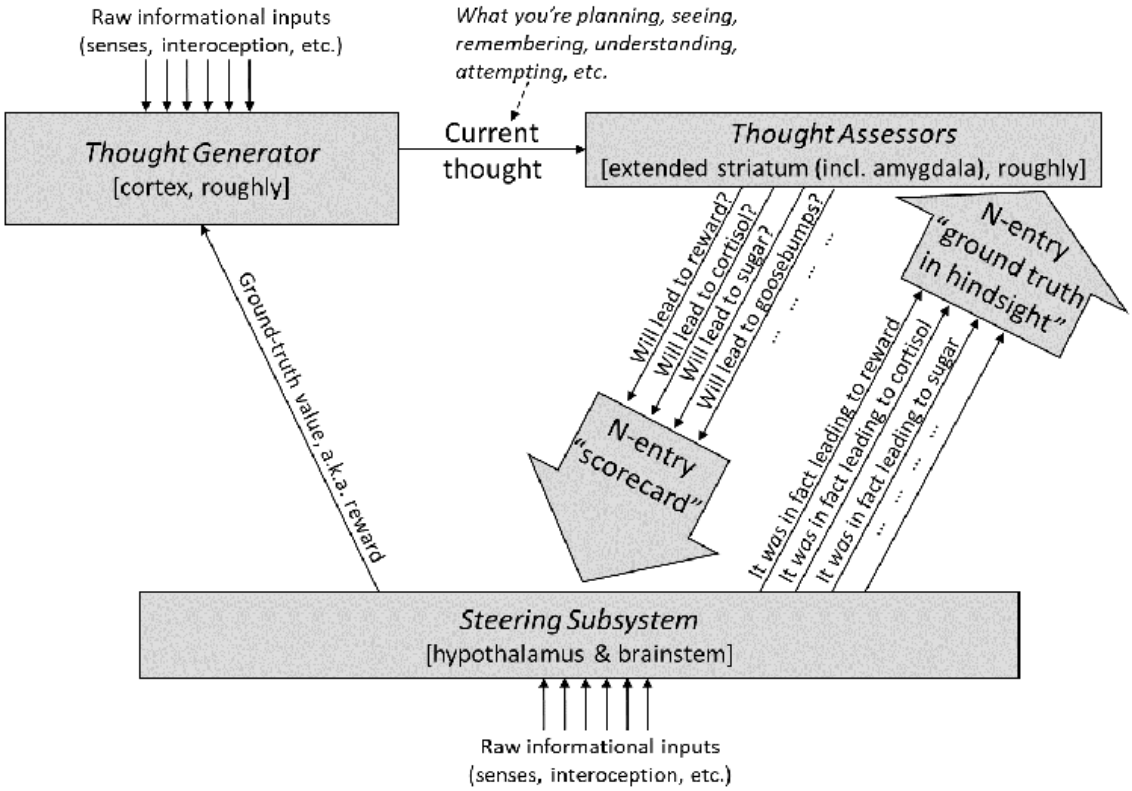
Table of contents:

- In Section 6.2, I'll present a big picture of motivation and decision-making in the human brain, and walk through how it works. The rest of the post will go through different parts of that picture in more detail. If you're in a hurry, I suggest reading to the end of Section 6.2 and then quitting.
- In Section 6.3, I'll talk about the so-called "Thought Generator", comprising (I think) the dorsolateral prefrontal cortex, sensory cortex, and other areas. (For ML readers familiar

with “actor-critic model-based RL”, the Thought Generator is more-or-less a combination of the “actor” and the “model”.) I’ll talk about the inputs and outputs of this module, and briefly sketch how its algorithm relates to neuroanatomy.
- In Section 6.4, I’ll talk about how values and rewards work in this picture, including the reward signal that drives learning and decision-making in the Thought Generator.
- In Section 6.5, I’ll go into a bit more detail about how and why thinking and decision-making needs to involve not only simultaneous comparisons (i.e., a mechanism for generating different options in parallel and selecting the most promising one), but also sequential comparisons (i.e., thinking of something, then thinking of something else, and comparing those two thoughts). For example, you might think: “Hmm, I think I’ll go to the gym. Actually, what if I went to the café instead?”
- In Section 6.6, I’ll comment on the common misconception that the Learning Subsystem is the home of ego-syntonic, internalized “deep desires”, whereas the Steering Subsystem is the home of ego-dystonic, externalized “primal urges”. I will advocate more generally against thinking of the two subsystems as two agents in competition; a better mental model is that the two subsystems are two interconnected gears in a single machine.

6.2 Big picture

Yes, this is literally a big picture, unless you’re reading on your cell phone. You saw a chunk of it in the previous post (Section 5.4), but now there are a few more pieces.

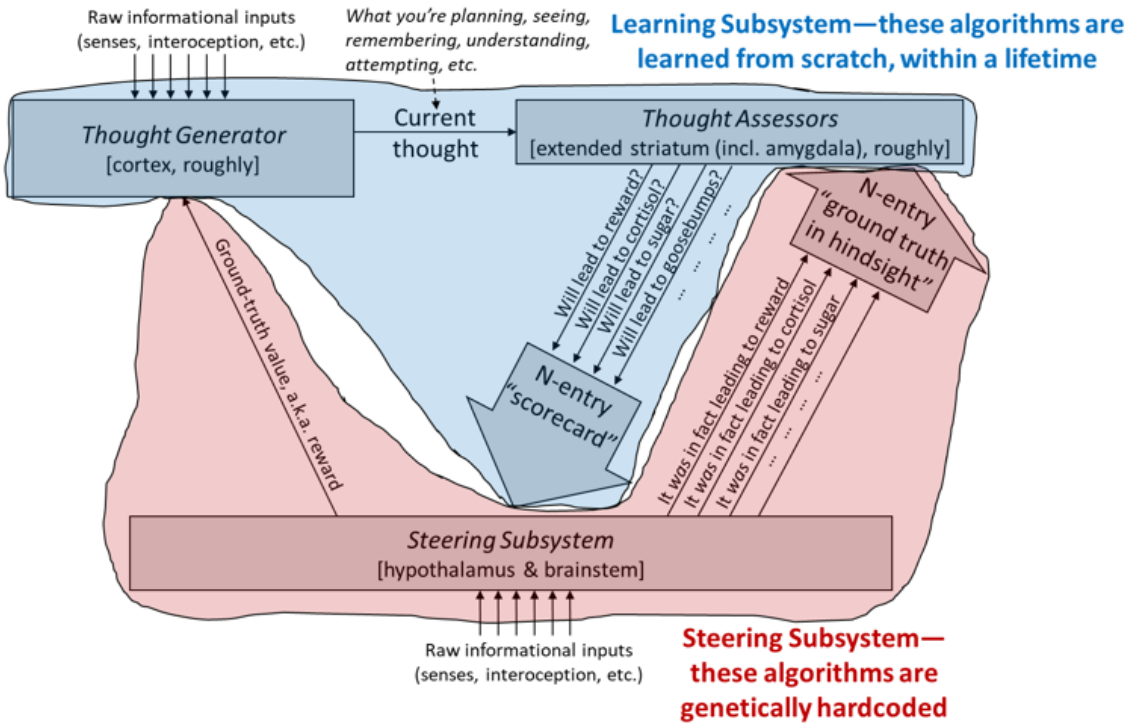


The big picture—The whole post will revolve around this diagram. Note that the bracketed neuroanatomy labels in the top two boxes are a bit provisional and certainly oversimplified. (dlPFC = dorsolateral prefrontal cortex; mPFC = medial prefrontal cortex.)

There’s a lot here, but don’t worry, I’ll walk through it bit by bit.

6.2.1 Relation to “two subsystems”

Here’s how this diagram fits in with my “two subsystems” perspective, first discussed in Post #3:



Same as above, but the two subsystems are highlighted in different colors.

6.2.2 Quick run-through

Before getting bogged down in details later in the post, I’ll just talk through the diagram:

1. Thought Generator generates a thought: The Thought Generator settles on a “thought”, out of the high-dimensional space of *every thought you can possibly think* at that moment. Note that this space of possibilities, while vast, is constrained by current sensory input, past sensory input, and everything else in your learned world-model. For example, if you’re sitting at a desk in Boston, it’s generally not possible for you to *think* that you’re scuba-diving off the coast of Madagascar. But you *can* make a plan, or whistle a tune, or recall a memory, or reflect on the meaning of life, etc.

2. Thought Assessors distill the thought into a “scorecard”: The Thought Assessors are a set of perhaps hundreds or thousands of “short-term predictor” circuits (Post #4), which I discussed more specifically in the previous post (#5). Each predictor is trained to predict a different signal from the Steering Subsystem. From the perspective of a Thought Assessor, everything in the Thought Generator (not just outputs but also latent variables) is *context*—information that they can use to make better predictions. Thus, if I’m thinking the thought “I’m going to eat candy right now”, a thought-assessor can predict “high probability of tasting something sweet very soon”, based *purely on the thought*—it doesn’t *need* to rely on either external behavior or sensory inputs, although those can be relevant context too.

3. The “scorecard” solves the interface problem between a learned-from-scratch world model and genetically-hardwired circuitry: Remember, the current thought and situation is

an insanely complicated object in a high-dimensional learned-from-scratch space of "all possible thoughts you can think". Yet we need the relatively simple, genetically-hardwired circuitry of the Steering Subsystem to analyze the current thought, including issuing a judgment of whether the thought is high-value or low-value (see Section 6.4 below), and whether the thought calls for cortisol release or goosebumps or pupil-dilation, etc. The "scorecard" solves that interfacing problem! It distills any possible thought / belief / plan / etc. into a genetically-standardized form that can be plugged directly into genetically-hardcoded circuitry.

4. The Steering Subsystem runs some genetically-hardwired algorithm: Its inputs are (1) the scorecard from the previous step and (2) various other information sources—pain, metabolic status, etc., all coming from its own brainstem sensory-processing system (see Post #3, Section 3.2.1). Its outputs include emitting hormones, motor commands, etc., as well as sending the "ground truth" supervisory signals shown in the diagram.[1]

5. The Thought Generator keeps or discards thoughts based on whether the Steering Subsystem likes them: More specifically, there's a ground-truth value (a.k.a. reward, yes I know those don't sound synonymous, see Post #5, Section 5.3.1). When the value is very positive, the current thought gets "strengthened", sticks around, and can start controlling behavior and summoning follow-up thoughts, whereas when the value is very negative, the current thought gets immediately discarded, and the Thought Generator summons a new thought instead.

6. Both the Thought Generator and the Thought Assessor "learn from scratch" over the course of a lifetime, thanks in part to these supervisory signals from the Steering Subsystem. Specifically, the Thought Assessors learn to make better and better predictions of their "ground truth in hindsight" signal (a form of Supervised Learning—see Post #4), while the Thought Generator learns to disproportionately generate high-value thoughts. (The Thought Generator learning-from-scratch process *also* involves predictive learning of sensory inputs—Post #4, Section 4.7.)

6.3 The "Thought Generator"

6.3.1 Overview

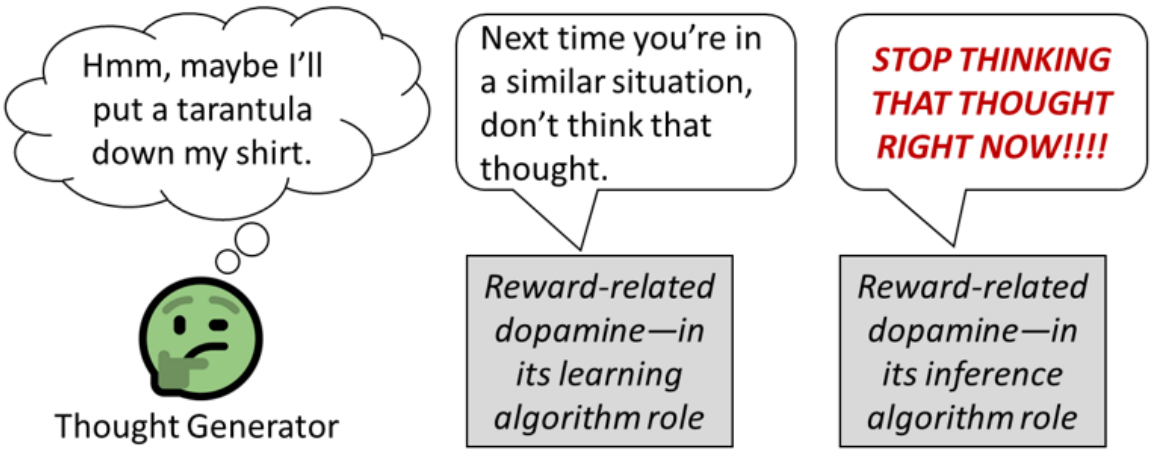
Go back to the big-picture diagram at the top. At the top-left, we find the Thought Generator. In terms of actor-critic model-based RL, the Thought Generator is roughly a combination of "actor" + "model", but not "critic". ("Critic" was discussed in the previous post, and more on it below.)

At our somewhat-oversimplified level of analysis, we can think of the "thoughts" generated by the Thought Generator as a combination of *constraints* (from predictive learning of sensory inputs) and *choices* (guided by reinforcement learning). In more detail:

- *Constraints* on the Thought Generator come from sensory input information, and ultimately from predictive learning of sensory inputs (Post #4, Section 4.7). For example, I *cannot* think the thought: *There is a cat on my desk and I'm looking at it right now.* There is no such cat, regrettably, and I can't just will myself to see something that obviously isn't there. I can *imagine* seeing it, but that's not the same thought.
- But *within* those constraints, there's more than one possible thought my brain can think at any given time. It can call up a memory, it can ponder the meaning of life, it can zone out, it can issue a command to stand up, etc. I claim that these "choices" are decided by a reinforcement learning (RL) system. This RL system is one of the main topics of this post.

6.3.2 Thought Generator inputs

The Thought Generator has a number of inputs, including sensory inputs and hyperparameter-shifting neuromodulators. But the main one of interest for this post is ground-truth value, a.k.a. reward. I'll talk about that in more detail later, but we can think of it as *an estimate of whether a thought is good or bad*, operationalized as "worth sticking with and pursuing" versus "deserving to be discarded so we can re-roll for a new thought". This signal is important *both* for learning to think better thoughts in the future, *and* for thinking good thoughts right now:



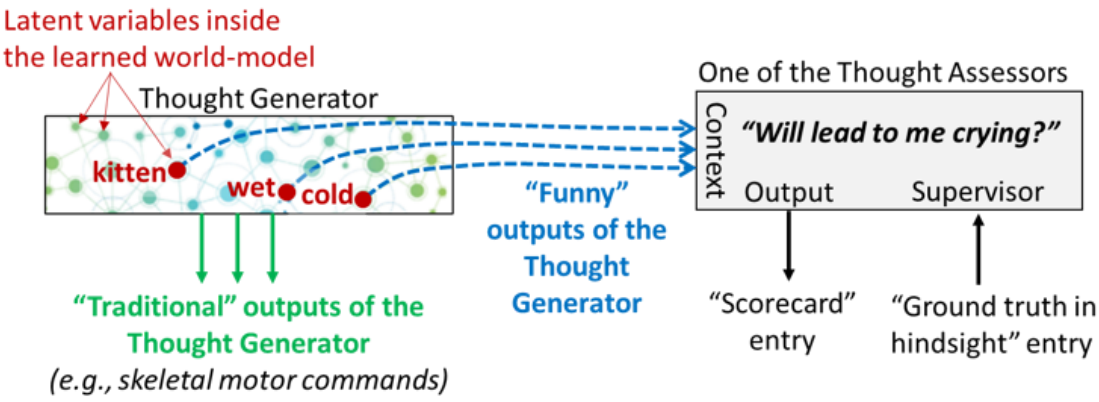
6.3.3 Thought Generator outputs

There are meanwhile a lot of signals going *out* of the Thought Generator. Some are what we intuitively think of as "outputs"—e.g., skeletal motor commands. Other outgoing signals are, well, a bit funny…

Recall the idea of "context" from Section 4.3 of Post #4: The Thought Assessors are short-term predictors, and a short-term predictor can in principle grab any signal in the brain and leverage it to improve its ability to predict its target signal. So if the Thought Generator has a world-model, then *somewhere* in the world-model is a configuration of latent variable activations that encode the concept "baby kittens shivering in the cold rain". We wouldn't normally think of those as "output signals"—I just said in the last sentence that they're latent variables! But as it happens, the "will lead to crying" Thought Assessor has grabbed a copy of those latent variables to use as context signals, and gradually learned through experience that these particular signals are strong predictors of me crying.

Now, as an adult, these "baby kittens in the cold rain" neurons in my Thought Generator are living a double-life:

- They are latent variables in my world-model—i.e., they and their web of connections will help me parse an image of baby kittens in the rain, if I see one, and to reason about what would happen to them, etc.
- Activating these neurons, e.g. via imagination, is *a way for me to call up tears on command*.

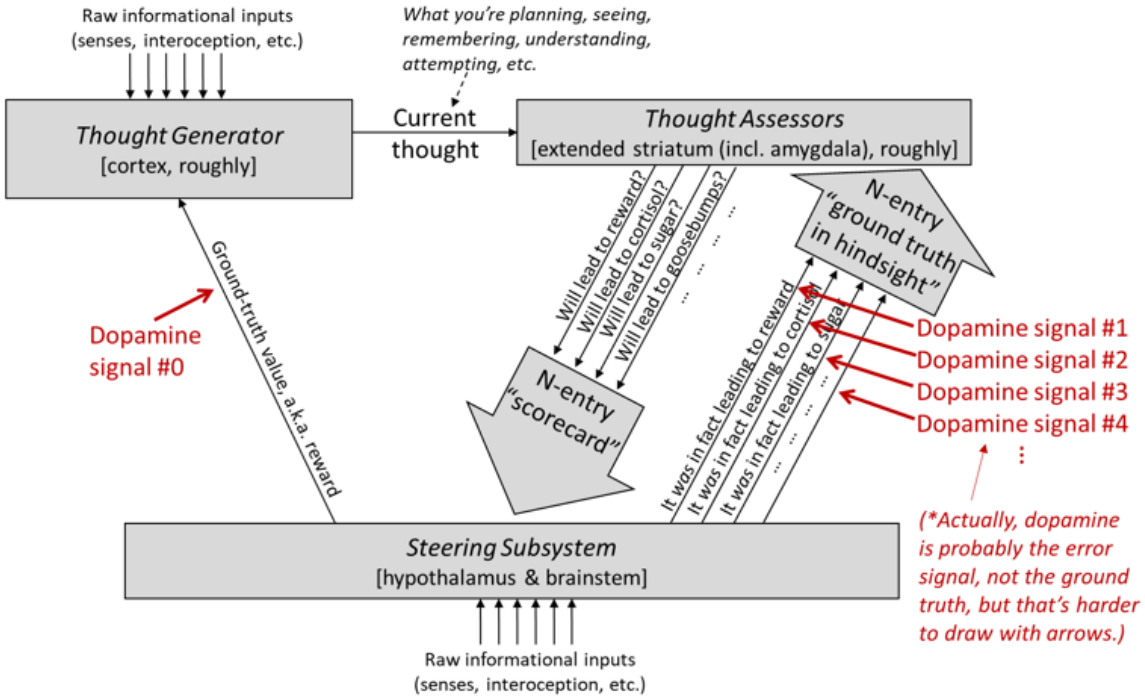


The Thought Generator (top left) has two types of outputs: the "traditional" outputs associated with voluntary behavior (green arrows) and the "funny" outputs wherein even latent variables in the model can directly impact involuntary behaviors (blue arrows).

6.3.4 Thought Generator neuroanatomy sketch

AUTHOR'S NOTE: When I first published this blog post, this section contained a discussion and diagrams of cortico-basal ganglia-thalamo-cortical loops, but it was very speculative and turned out to be wrong in various ways. It's not too relevant for the series anyway, so I'm deleting it. I'll write a corrected version at some point. Sorry!

Here's the updated dopamine diagram from the previous post:



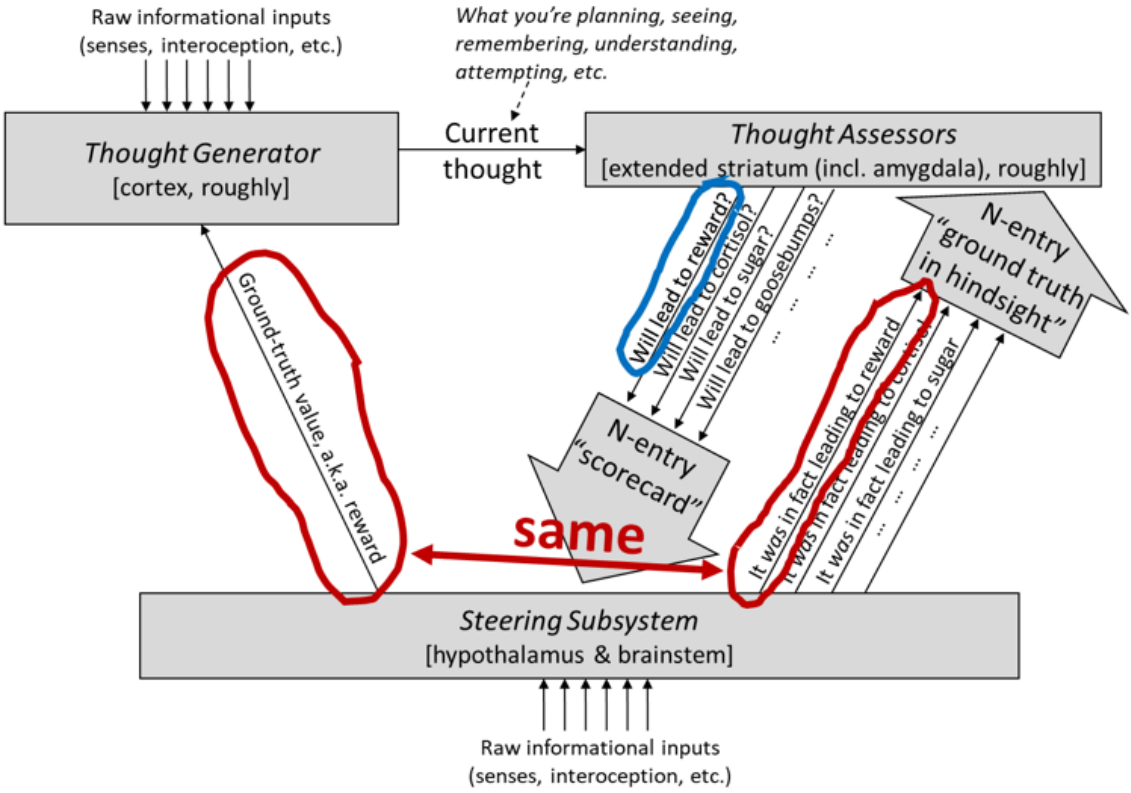
The "mesolimbic" dopamine signals on the right were discussed in the previous post (Section 5.5.6). The "mesocortical" dopamine signal on the left is new to this post. (I think there are *even more* dopamine signals in the brain, not shown here. They're off-topic for this series, but see discussion here.)

There are many more implementation details inside the Thought Generator that I'm not discussing. However, this bare-bones section is more-or-less sufficient for my forthcoming posts on AGI safety. The gory details of the Thought Generator, like the gory details of almost everything else in the Learning Subsystem, are mainly helpful for *building* AGI.

6.4 Values and rewards

6.4.1 The cortex proposes a "value" estimate, but the Steering Subsystem may choose to override

There are two "values" in the diagram (it looks like three, but the two red ones are the same):



Two types of "value" in my model

The blue-circled signal is the value estimate from the corresponding Thought Assessor in the cortex. The red-circled signal (again, it's one signal drawn twice) is the corresponding "ground truth" for what the value estimate should have been. (Recall that "ground-truth value" is a synonym for "reward"; yes I know that sounds wrong, see previous post (Section 5.3.1) for discussion.)

Just like the other "long-term predictors" discussed in the previous post, the Steering Subsystem can choose between "defer-to-predictor mode" and "override mode". In the former, it sets the red equal to the blue, as if to say "OK, Thought Assessor, sure, I'll take your word for it". In the latter, it ignores the Thought Assessor's proposal, and its own internal circuitry outputs some different value.[2]

Why might the Steering Subsystem override the Thought Assessor's value estimate? Two factors:

- First, the Steering Subsystem might be acting on information from *other* (non-value) Thought Assessors. For example, in the Dead Sea Salt Experiment (see previous post, Section 5.5.5), the value estimator says "bad things are going to happen", but meanwhile the Steering Subsystem is getting an "I'm about to taste salt" prediction in the context of a state of salt-deprivation. So the Steering Subsystem says to itself "Whatever is happening now is very promising; the value estimator doesn't know what it's talking about!"
- Second, the Steering Subsystem might be acting on its own information sources, independent of the Learning Subsystem. In particular, the Steering Subsystem has its own sensory-processing system (see Post #3, Section 3.2.1), which can sense biologically-relevant cues like pain status, hunger status, taste inputs, the sight of a slithering snake, the smell of a potential mate, and so on. All these things and more can be possible bases for overruling the Thought Assessor, i.e., setting the red-circled signal to a different value than the blue-circled one.

Interestingly (and unlike in textbook RL), in the big picture, *the blue-circled signal doesn't have a special role in the algorithm*, as compared to the other Thought Assessors. It's just one of many inputs to the Steering Subsystem's hardwired algorithm for deciding what to put into the red-circled signal. The blue-circled signal might be an *especially important* signal in practice, weighed more heavily than the others, but ultimately everything is in the same pot. In fact, my longtime readers will recall that last year I was writing posts that *omitted* the blue-circled value signal from the list of Thought Assessors! I now think that was a mistake, but I retain a bit of that same attitude.

6.5 Decisions involve not only simultaneous but also sequential comparisons of value

Here's a "simultaneous" model of decision-making, as described by *The Hungry Brain* by Stephan Guyenet in the context of studies on lamprey fish:

> Each region of the pallium [= lamprey equivalent of cortex] sends a connection to a particular region of the striatum, which (via other parts of the basal ganglia) returns a connection back to the same starting location in the pallium. This means that each region of the pallium is reciprocally connected with the striatum via a specific loop that regulates a particular action…. For example, there's a loop for tracking prey, a loop for fleeing predators, a loop for anchoring to a rock, and so on. Each region of the pallium is constantly whispering to the striatum to let it trigger its behavior, and the striatum always says "no!" by default. In the appropriate situation, the region's whisper becomes a shout, and the striatum allows it to use the muscles to execute its action.

I endorse this as *part* of my model of decision-making, but only part of it. Specifically, this is one of the things that's happening when the Thought Generator generates a thought. Indeed, my diagram in Section 6.3.4 above takes obvious inspiration from the model above. Different *simultaneous* possibilities are being compared.

The other part of my model is comparisons of *sequential* thoughts. You think a thought, and then you think a different thought (possibly very different, or possibly a refinement of the first thought), and the two are implicitly compared (by the Steering Subsystem picking a ground-truth value based on the temporal dynamics of Thought Assessors jumping up and down, for example), and if the second thought is worse, it gets weakened such that a new thought can replace it (and the new thought might be the first thought re-establishing itself).

I could cite experiments for the sequential-comparison aspect of decision-making (e.g. Figure 5 of this paper, which is arguing the same point as I am), but do I really need to? Introspectively, it's obvious! You think: "Hmm, I think I'll go to the gym. Actually, what if I went to the café instead?" You're imagining one thing, and then another thing.

And I don't think this is is a humans-vs-lampreys thing. My hunch is that comparisons of sequential thoughts is universal in vertebrates. As an illustration of what I mean:

6.5.1 Made-up example of what comparison-of-sequential-thoughts might look like in a simpler animal

Imagine a simple, ancient, little fish swimming along, navigating to the cave where it lives. It gets to a ~~fork in the road,~~ ummm, "fork in the kelp forest"? Its current navigation plan involves continuing left to its cave, but it also has the option of turning right to go to the reef, where it often forages.

Seeing this path to the right, I claim that its navigation algorithm reflexively loads up a plan: "I'm will turn right and go to the reef." Immediately, this new plan is evaluated and compared to the old plan. If the new plan seems worse than the old plan, then the new thought gets shut down, and the old thought ("I'm going to my cave") promptly reestablishes itself. The fish continues to its cave, as originally planned, without skipping a beat. Whereas if instead the new plan seems *better* than the old plan, then the new plan gets strengthened, sticks around, and orchestrates motor commands. And thus the fish turns to the right and goes to the reef instead.

(In reality, I don't know much about little ancient fish, but rats at a fork in the ~~road~~ maze are known to imagine both possible navigation plans in succession, based on measurements of hippocampus neurons—ref.)

6.5.2 Comparison-of-sequential-thoughts: why it's necessary

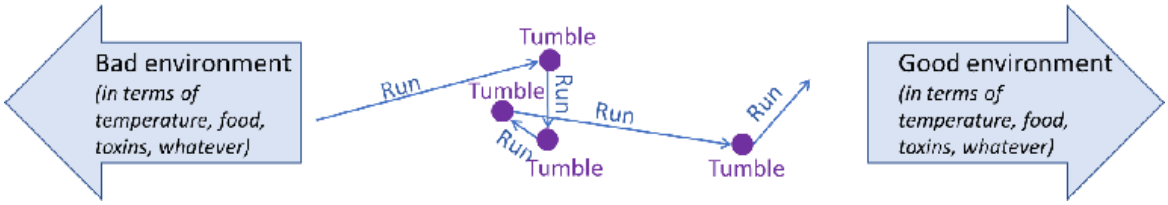
In my view, thoughts are complicated. To think the thought "I will go to the café", you're not just activating some tiny cluster of dedicated go-to-the-café neurons. Instead, it's a distributed pattern involving practically every part of the cortex. You can't simultaneously think "I will go to the café" and "I will go to the gym", because they would involve different activity patterns of the same pools of neurons. They would cross-talk. Thus, the only possibility is thinking the thoughts in sequence.

As a concrete example of what I have in mind, think of how a Hopfield network can't recall twelve different memories simultaneously. It has multiple stable states, but you can only explore them sequentially, one after the other. Or think about grid cells and place cells, etc.

6.5.3 Comparison-of-sequential-thoughts: how it might have evolved

From an evolutionary perspective, I imagine that comparison-of-sequential-thoughts is a distant descendent of a very simple mechanism akin to the run-and-tumble mechanism in swimming bacteria.

In the run-and-tumble mechanism, a bacterium swims in a straight line ("runs"), and periodically changes to a new random direction ("tumbles"). But the trick is: when the bacterium's situation / environment is getting *better*, it tumbles *less* frequently, and when it's getting *worse*, it tumbles *more* frequently. Thus, it winds up moving in a good direction (on average, over time).



Starting with a simple mechanism like that, one can imagine adding progressively more bells and whistles. The palette of behavioral options can get more and more complex, eventually culminating in "every thought you can possibly think". The methods of evaluating whether the current plan is good or bad can get faster and more accurate, eventually involving learning-algorithm-based predictors as in the previous post. The new behavioral options to tumble *into* can be picked via clever learning algorithms, rather than randomly. Thus, it seems to me that there's a smooth path all the way from something-akin-to-run-and-tumble to the intricate, finely-tuned, human brain system that I'm talking about in this series. (Other musings on run-and-tumble versus human motivation: 1, 2.)

6.6 Common misconceptions

6.6.1 The distinction between internalized ego-syntonic desires and externalized ego-dystonic urges is unrelated to Learning Subsystem vs. Steering Subsystem

(See also: my post "(Brainstem, Neocortex) ≠ (Base Motivations, Honorable Motivations)".)

Many people (including me) have a strong intuitive distinction between ego-syntonic drives that are "part of us" or "what we want", versus ego-dystonic drives that feel like urges which intrude upon us from the outside.

For example, a food snob might say "I love fine chocolate", while a dieter might say "I have an urge to eat fine chocolate".

6.6.1.1 The explanation I like

I would claim that these two people are basically describing the same feeling, with essentially the same neuroanatomical locations and essentially the same relation to low-level brain algorithms. But the food snob is *owning* that feeling, and the dieter is *externalizing* that feeling.

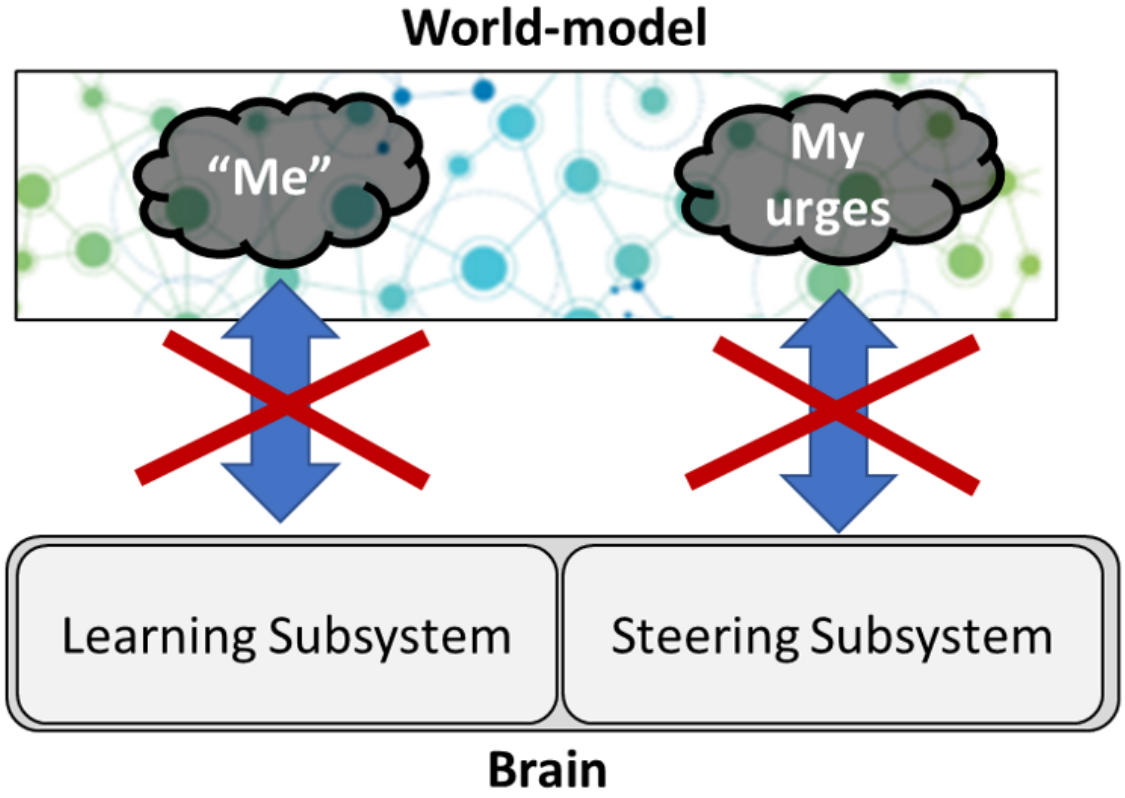
These two different self-concepts go hand-in-hand with two different "higher-order preferences": the food snob *wants to want to eat fine chocolate* while the dieter *wants to not want to eat fine chocolate*.

This leads us to a straightforward psychological explanation for why the food snob and dieter conceptualize their feelings differently:

- The food snob finds it *appealing* to think of “the desire I feel for fine chocolate” as “part of who I am”. So he does.
- The dieter finds it *aversive* to think of “the desire I feel for fine chocolate” as “part of who I am”. So he doesn’t.

6.6.1.2 The explanation I don’t like

Many people (including Jeff Hawkins, see Post #3) notice the distinction described above, and separately, they endorse the idea (as I do) that the brain has a Learning Subsystem and Steering Subsystem (again see Post [Intro to brain-like-AGI safety] 6. Big picture of motivation, decision-making, and RL



> Most people I talk to, including me, have separate concepts in our learned world-models for “me” and “my urges”. I claim that these concepts did *NOT* come out of veridical introspective access to our own neuroanatomy. And in particular, they do *not* correspond respectively to the Learning & Steering Subsystems.

I think this model is wrong. At the *very* least, if you want to endorse this model, then you need to reject approximately everything I’ve written in this and my previous four posts.

In *my* story, if you’re trying to abstain from chocolate, but also feel an urge to eat chocolate, then:

- You have an urge to eat chocolate because the Steering Subsystem approves of the thought “I am going to eat chocolate right now”; AND
- You’re trying to abstain from chocolate because the Steering Subsystem approves of the thought “I am abstaining from chocolate”.

(Why would the Steering Subsystem approve of the latter? It depends on the individual, but it’s probably a safe bet that social instincts are involved. I’ll talk more about social instincts in Post #13. If you want an example with less complicated baggage, imagine a lactose-intolerant person trying to resist the urge to eat yummy ice cream right now, because it will make them feel really sick later on. The Steering Subsystem likes plans that result in not feeling sick, and *also* likes plans that result in eating yummy ice cream.)

6.6.2 The Learning Subsystem and Steering Subsystem are not two agents

Relatedly, another frequent error is treating either the Learning Subsystem or Steering Subsystem by itself as a kind of independent agent. This is wrong on both sides:

- The Learning Subsystem cannot think any thoughts unless the Steering Subsystem has endorsed those thoughts as being worthy of being thunk.
- Meanwhile, the Steering Subsystem does not understand the world, or itself. It has no explicit goals for the future. It’s just a relatively simple, hardcoded input-output machine.

As an example, the following is *entirely possible*:

1. The Learning Subsystem generates the thought *“I am going to surgically alter my own Steering Subsystem”*.
2. The Thought Assessors distill that thought down to the “scorecard”.
3. The Steering Subsystem gets the scorecard and runs it through its hardcoded heuristics, and the result is: “Very good thought, go right ahead and do it!”

Why not, right? I’ll talk more about that example in later posts.

If you just read the above example, and you’re thinking to yourself “Ah! This is a case where the Learning Subsystem has outwitted the Steering Subsystem”, then *you’re still not getting it*.

(Maybe instead try imagining the Learning Subsystem & Steering Subsystem as two interconnected gears in a single machine.)

1. ^

   As in the previous post, the term “ground truth” here is a bit misleading, because sometimes the Steering Subsystem will just *defer* to the Thought Assessors.

2. ^

   As in the previous post, I don’t *really* believe there is a pure dichotomy between “defer-to-predictor mode” and “override mode”. In reality, I’d bet that the Steering Subsystem can partly-but-not-entirely defer to the Thought Assessor, e.g. by taking a weighted average between the Thought Assessor and some other independent calculation.

[Intro to brain-like-AGI safety] 7. From hardcoded drives to foresighted plans: A worked example

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

Part of the "Intro to brain-like-AGI safety" post series .

(This post substantially overlaps with my post from last August, Value loading in the human brain: a worked example . Compared to that older post, this version has numerous minor edits for clarity, correctness, and fitting-into-the-flow-of-this-series.)

7.1 Post summary / Table of contents

The previous post presented a big picture of how I think motivation works in the human brain, but it was a bit abstract. In this post, I will walk through an example. To summarize, the steps will be:

1. (Section 7.3) Our brains gradually develop a probabilistic generative model of the world and ourselves;
2. (Section 7.4) There's a "credit assignment" process, where something in the world-model gets flagged as "good";
3. (Section 7.5) There's a reward prediction error signal roughly related to the *time-derivative of the expected probability of the "good" thing*. This signal drives us to "try" to make the "good" thing happen, including via foresighted planning.

All human goals and motivations come *ultimately* from relatively simple, genetically-hardcoded circuits in the Steering Subsystem (hypothalamus and brainstem), but the details can be convoluted in some cases. For example, sometimes I'm motivated to do a silly dance in front of a full-length mirror. Exactly what genetically-hardcoded hypothalamus or brainstem circuits are upstream of that motivation? I don't know! Indeed, I claim that the answer is currently Not Known To Science. I think it would be well worth figuring out! Umm, well, OK, maybe that *specific* example is not worth figuring out. But the broader project of reverse-engineering certain aspects of the human Steering Subsystem (see my discussion of "Category B" in Post #3)—especially those upstream of social instincts like altruism and status-drive—is a project that I consider desperately important for AGI safety, and utterly neglected. More on that in Posts #12–#13.

In the meantime, I'll pick an example of a goal that to a first approximation comes from an *especially straightforward and legible* set of Steering Subsystem circuitry. Here goes.

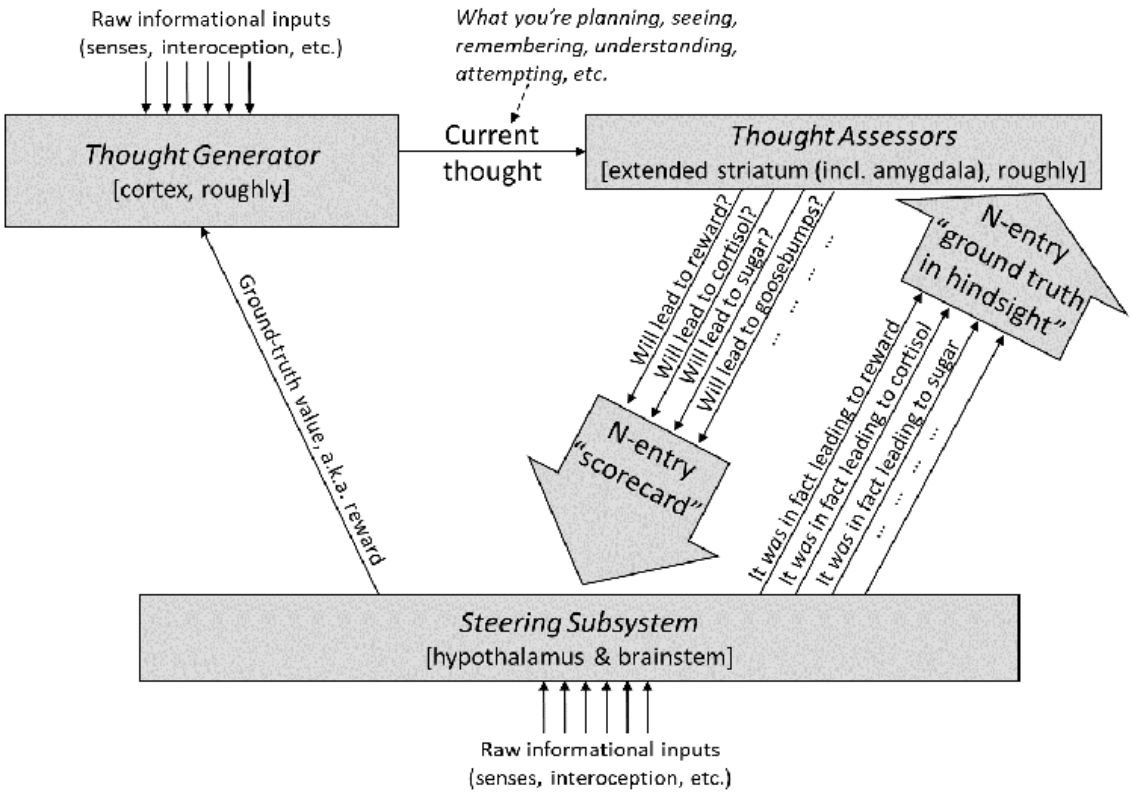
Let's say (purely hypothetically… 👀) that I ate a slice of prinsesstårta cake two years ago, and it was really yummy, and ever since then I've wanted to eat one again. ***So my running example of an explicit goal in this post will be "I want a slice of prinsesstårta".***

Prinsesstårta cake. I suggest eating some, in order to better understand this blog post. For science! (Image source: my favorite local bakery.)

Eating a slice of prinsesstårta is not my only goal in life, or even a particularly important one—so it has to trade off against my other goals and desires—but it is nevertheless *a* goal of mine (at least when I'm thinking about it), and I would indeed make complicated plans to try bring about that goal. Like, for example, dropping subtle hints to my family. In blog posts. When my birthday is coming up. Purely hypothetically!!

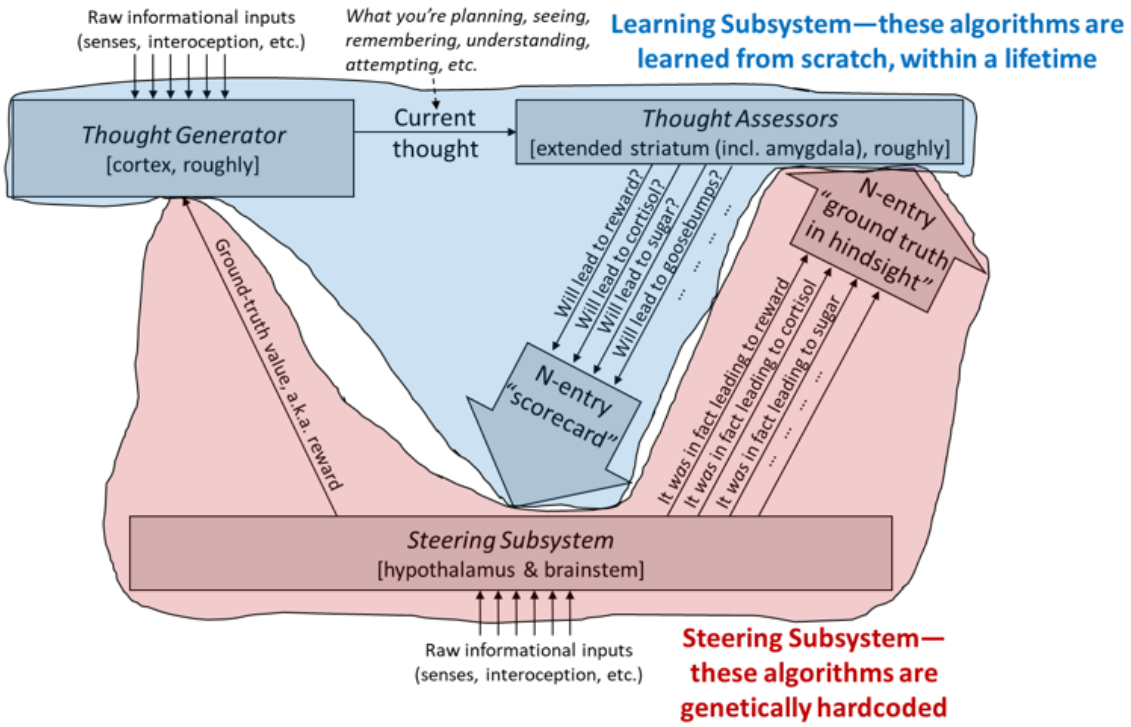
7.2 Reminder from the previous post: big picture of motivation and decision-making

From the previous post, here's my diagram of motivation in the brain:



See previous post for details.

As also discussed in the previous post, we can split this up by which parts are "hardcoded" by the genome, versus learned within a lifetime—i.e., Steering Subsystem versus Learning Subsystem:



7.3 Building a probabilistic generative world-model in the cortex

The first step in our story is that, over my lifetime, my cortex (specifically, the Thought Generator in the top-left of the diagram above) has been building up a probabilistic generative model, mostly by predictive learning of sensory inputs (Post #4, Section 4.7) (a.k.a. "self-supervised learning").

Basically, we learn patterns in our sensory input, and patterns in the patterns, etc., until we have a nice predictive model of the world (and of ourselves)—a giant web of interconnected entries like "grass" and "standing up" and "slices of prinsesstårta".

Predictive learning of sensory inputs is not fundamentally dependent on supervisory signals from the Steering Subsystem. Instead, "the world" provides the ground truth about whether a prediction was correct. Contrast this with, for example, navigating the tradeoff between searching-for-food versus searching-for-a-mate: there is no "ground truth" in the environment for whether the animal is trading off optimally, except after generations of hindsight. In that case, we *do* need supervisory signals from the Steering Subsystem, which estimate the "correct" tradeoff using heuristics hardcoded by evolution. You can kinda think of the is/ought divide, with the Steering Subsystem providing the "ought" ("to maximize genetic fitness, what ought the organism to do?") and predictive learning of sensory inputs providing the "is" ("what is likely to happen next, under such-and-such circumstances?") That said, the Steering Subsystem is *indirectly* involved even in predictive learning of sensory inputs—for example, I can be motivated to go learn about a topic.

Anyway, every thought I can possibly think, and every plan I can possibly plan, can be represented as some configuration of this generative world-model data structure. The data structure is also continually getting edited, as I learn and experience new things.

When you think of this world-model data structure, imagine many terabytes of inscrutable entries—imagine things like, for example,

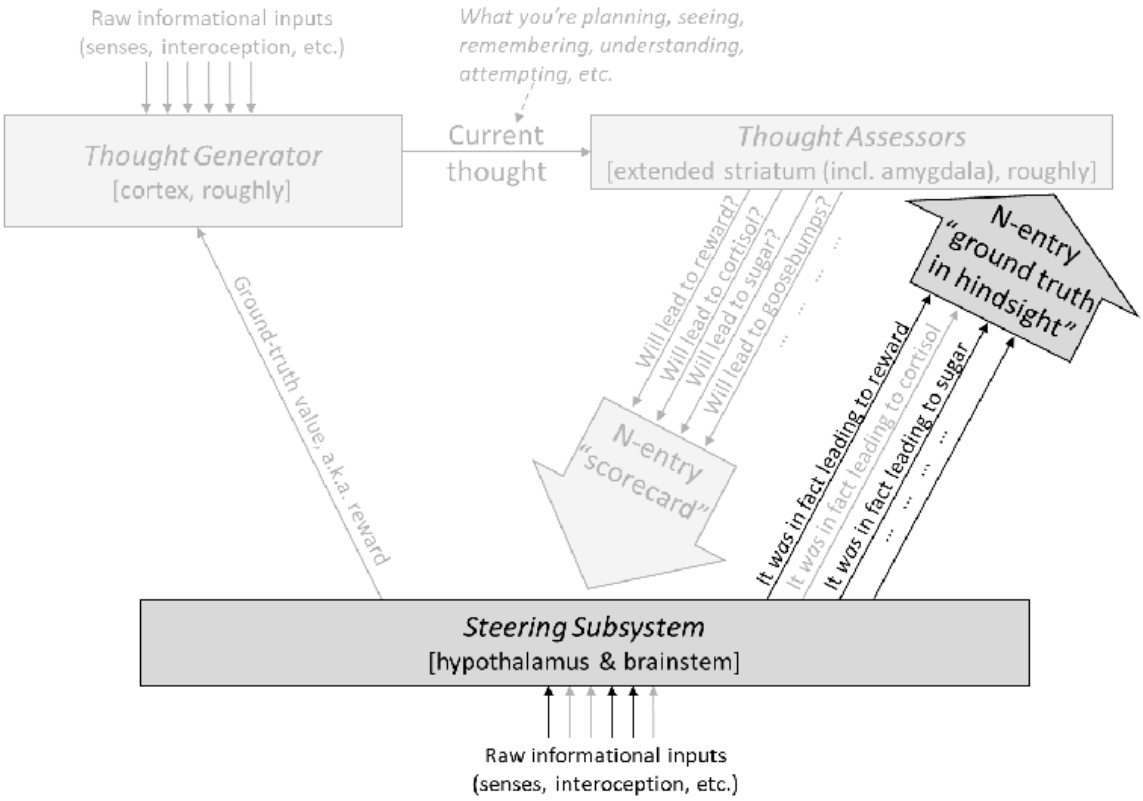
"PATTERN 847836 is defined as the following sequence: {PATTERN 278561, then PATTERN 657862, then PATTERN 128669}."

Some entries have references to sensory inputs and/or motor outputs. And that giant inscrutable mess comprises my entire understanding of the world and myself.

7.4 Credit assignment when I first bite into the cake

As I mentioned at the top, on a fateful day two years ago, I ate a slice of prinsesstårta, and it was *really good*.

Step back to a couple seconds earlier, as I was bringing the cake towards my mouth to take my first-ever bite. At that moment, I didn't yet have any particularly strong expectation of what it would taste like, or how it would make me feel. But once it was in my mouth, mmmmmmm, oh wow, that's good cake.



Relevant parts of the diagram for what happened when I took my first surprisingly-delicic bite of prinsesstårta, two years ago.

So, as I took that bite, my body had a suite of autonomic reactions—releasing certain hormones, salivating, changing my heart rate and blood pressure, etc. Why? The key is that, as described in Post #3, Section 3.2.1, all sensory inputs split:

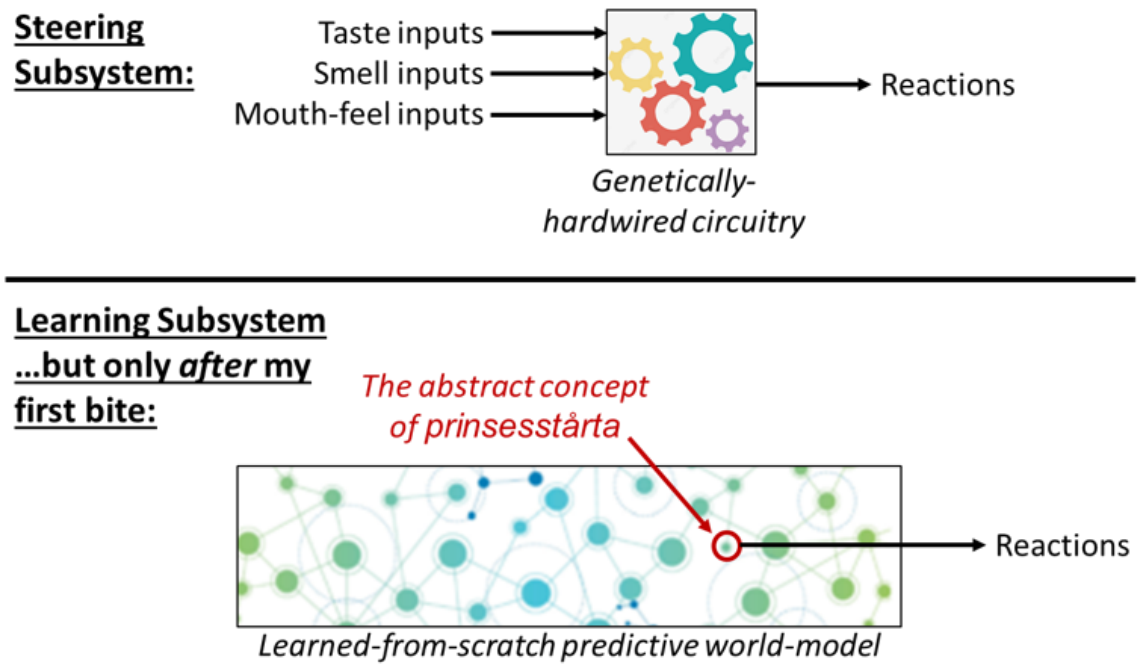
- One copy of any given sensory signal goes to the Learning Subsystem, to be integrated into the predictive world-model. (See “Informational inputs” at the top left of the diagram.)
- A second copy of the same signal goes to the Steering Subsystem, where it serves as an input to genetically-hardwired circuitry. (See “Informational inputs” at the bottom-center of the diagram.)

Taste bud inputs are no exception: the former signal winds up at the gustatory cortex within the insula (part of the neocortex, in the Learning Subsystem), the latter at the gustatory nucleus of the medulla (part of the brainstem, in the Steering Subsystem). After its arrival at the medulla, the taste inputs feed into various genetically-hardcoded brainstem circuits, which, when also prompted with the taste and mouth-feel of the cake, and also accounting for my current physiological state and so on, execute all those autonomic reactions I mentioned.

As I mentioned, *before* I first bit into the cake, I didn’t expect it to be that good. Well, maybe *intellectually* I expected it—if you had asked me, I would have *said* and *believed* that the cake would be really good. But I didn’t *viscerally* expect it.

What do I mean by “viscerally”? What’s the difference? The things I *viscerally* expect are over on the “Thought Assessor” side. People don’t have voluntary control over their Thought Assessors—the latter are trained exclusively by the “ground truth in hindsight” signals from the brainstem. You do have *some* ability to manipulate them by controlling what you’re thinking about, as discussed in the previous post (Section 6.3.3), but to a first approximation they’re doing their own thing, independent of what you *want* them to be doing. From an evolutionary perspective, this design makes good sense as a defense against wireheading—see my post Reward Is Not Enough.

So when I bit into the cake, my Thought Assessors were wrong! They expected the cake to cause *mild* “yummy”-related autonomic reactions, but in fact the cake caused *intense* “yummy”-related autonomic reactions. And the Steering Subsystem *knew* that the Thought Assessors had been wrong. So it sent correction signals up to the Thought Assessor algorithms, as shown in the diagram above. Those algorithms then edited themselves, so that going forward, every time I bring a fork-full of prinsesstårta towards my mouth, the Thought Assessors will be more liable to predict intense hormones, goosebumps, reward, and all the other reactions that I did in fact get.



A cool thing just happened here. We started with a simple-ish hardwired algorithm: Steering Subsystem circuits turning certain types of taste inputs into certain hormones and autonomic reactions. But then we transferred that information into *functions on the learned world-model*—recall that giant inscrutable database I was talking about in the previous section.

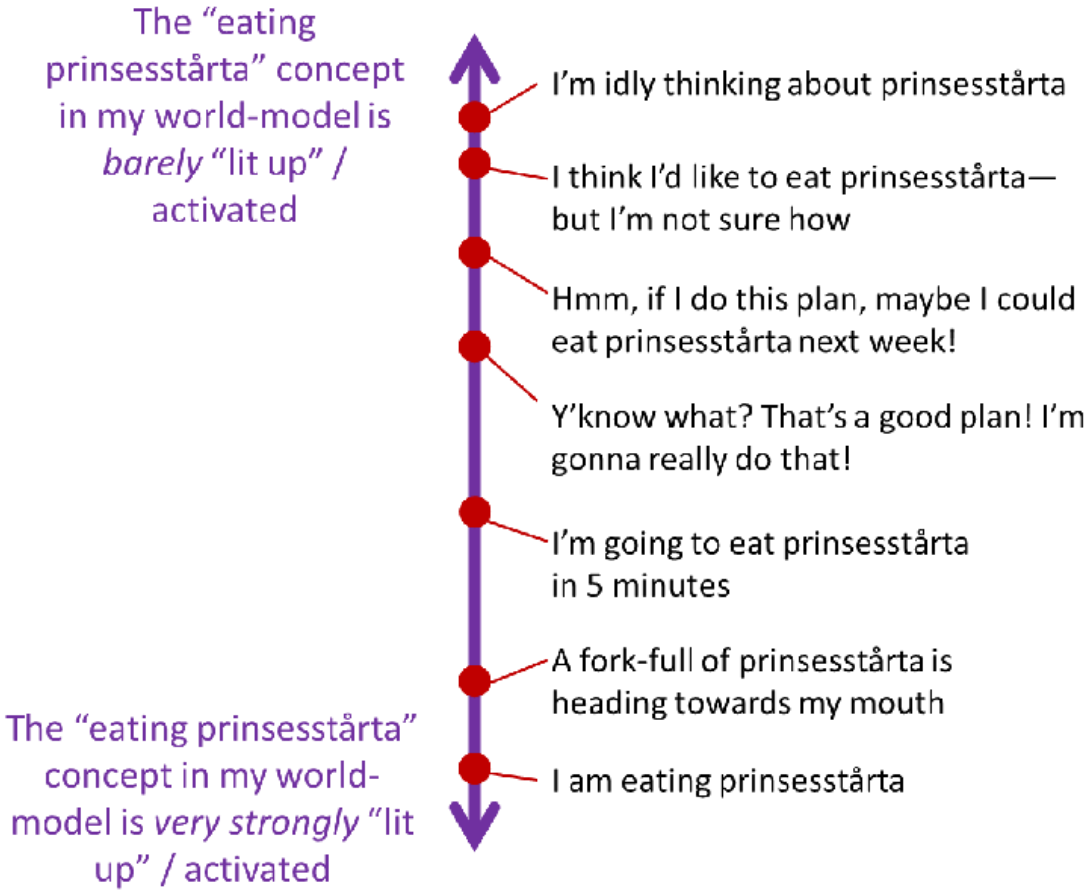
(Let me pause to spell this out a bit: The "ground truth in hindsight" signal tweaks some of the Thought Assessors. The Thought Assessors, you'll recall from Post #5, are a set of maybe hundreds of models, each trained by supervised learning. The inputs to those trained models, or what I call "context" signals (see Post #4), include neurons from inside the predictive world-model that encode "what thought is being thunk right now". So we wind up with a function (trained model) whose input includes things like "whether my current thought activates the abstract concept of prinsesstårta", and whose output is a signal that tells the Steering Subsystem to consider salivating etc.)

I call this step—where we edit the Thought Assessors—"credit assignment". Much more about that process in upcoming posts, including how it can go awry.

So now the Thought Assessors have learned that whenever the "myself eating prinsesstårta" concept "lights up" in the world-model, they should issue predictions of the corresponding hormones, other reactions, and reward.

7.5 Planning towards goals via reward-shaping

I don't have a particularly rigorous model for this step, but I think I can lean on intuitions a bit, in order to fill in the rest of the story:



Remember, ever since my first bite of prinsesstårta two years ago, the Thought Assessors in my brain have been inspecting each thought I think, checking whether the "myself eating prinsesstårta" concept in my world-model is "lit up" / "activated", and to the extent that it is, issuing a suggestion to prepare for rewards, salivation, goosebumps, and so on.

The diagram above suggests a series of thoughts that I think would "light up" the world-model concept more and more, as we go from top to bottom.

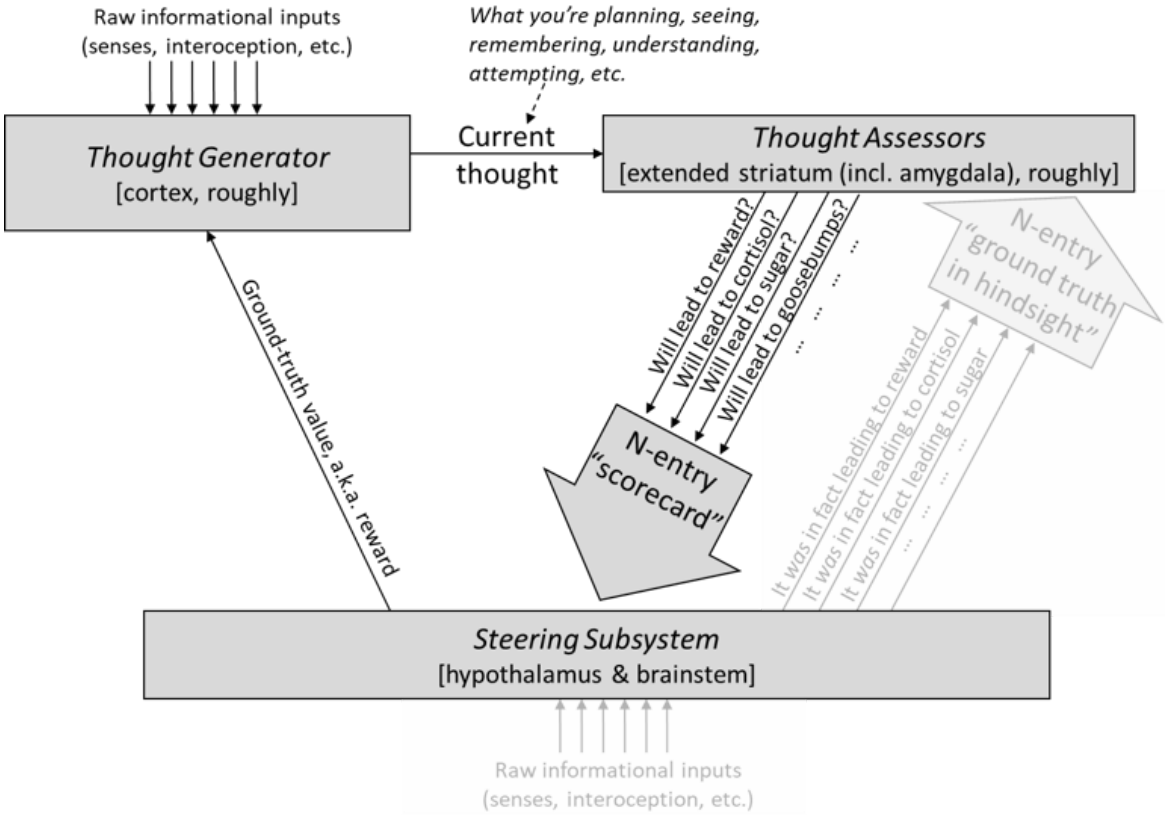
To get the intuition here, maybe try replacing "prinsesstårta" with "super-salty cracker". Then go down the list, and try to feel how each thought would make you salivate more and more. Or better yet, replace "eating prinsesstårta" with "asking my crush out on a date", go down the list, and try to feel how each thought makes your heart rate jump up higher and higher.

Here's another way to think about it: If you imagine the world-model being vaguely like a PGM, you can imagine that the "degree of pattern-matching" corresponds roughly to the probability assigned to the "eating prinsesstårta" node in the PGM. For example, if you're confident in X, and X weakly implies Y, and Y weakly implies Z, and Z weakly implies "eating prinsesstårta", then "eating prinsesstårta" gets a very low but nonzero probability, a.k.a. weak activation, and this is akin to having a far-fetched but not completely impossible plan to eat prinsesstårta. (Don't take this paragraph too literally, I'm just trying to summon intuitions here.)

I'm really hoping this kind of thing is intuitive. After all, I've seen it reinvented numerous times! For example, David Hume: "The first circumstance, that strikes my eye, is the great resemblance betwixt our impressions and ideas in every other particular, except their degree

of force and vivacity." And here's William James: "It is hardly possible to confound the liveliest image of fancy with the weakest real sensation." In both these cases, I think the authors are gesturing at the idea that imagination activates some of the same mental constructs (latent variables in the world-model) as perception does, but that imagination activates them more weakly than perception.

OK, if you're still with me, let's go back to my decision-making model, now with different parts highlighted:



Relevant parts of the diagram for the process of making and executing a foresighted plar to procure prinsesstårta.

Again, every time I think a thought, the Steering Subsystem looks at the corresponding "scorecard", and issues a corresponding reward. Recall also that the active thought / plan gets thrown out when its reward signal is negative, and it gets kept and strengthened when its reward is positive.

I'll oversimplify for a second, and ignore everything except the value function (a.k.a. The "Will lead to reward" Thought Assessor). And I'll also assume the Steering Subsystem just defers to that proposed value, rather than overruling it (see Post #6, Section 6.4.1). In this case, each time our thoughts move down a notch on the purple arrow diagram above—from idle musing about prinsesstårta, to a hypothetical plan to get prinsesstårta, to a decision to get prinsesstårta, etc.—there's an *immediate positive* reward, so that the new thought gets strengthened, and gets to establish itself. And conversely, each time we move back *up* the list—from decision to hypothetical plan to to idle musing—there's an *immediate negative* reward, so that thought gets thrown out and we go back to whatever we were thinking before. It's a ratchet! The system naturally pushes its way down the list, making and executing a good plan to eat cake.

So there you have it! From this kind of setup, I think we're well on the way to explaining the full suite of behaviors associated with humans doing foresighted planning towards explicit goals—including knowing that you have the goal, making a plan, pursuing instrumental strategies as part of the plan, replacing good plans with even better plans, updating plans as the situation changes, pining in vain for unattainable goals, and so on.

7.5.1 The other Thought Assessors. Or: The heroic feat of ordering a cake for next week, when you're feeling nauseous right now

By the way, what of the *other* Thought Assessors? Prinsesstårta, after all, is not *just* associated with "will lead to rewards", but also "will lead to sweet taste", "will lead to salivation", etc. Do those play any role?

Sure! For one thing, as I bring the fork towards my mouth, on the verge of consummating my cake-eating plan, I'll start salivating and releasing cortisol in preparation.

But what about the process of foresighted planning (calling the bakery etc.)? I think the other, non-value-function, Thought Assessors are relevant there too—at least to some extent.[1]

For example, imagine you're feeling terribly nauseous. Of course your Steering Subsystem *knows* that you're feeling terribly nauseous. And then suppose it sees you thinking a thought that seems to be leading towards eating. In that case, the Steering Subsystem may say: "That's a terrible thought! Negative reward!"

OK, so you're feeling nauseous, and you pick up the phone to place your order at the bakery. This thought gets weakly but noticeably flagged by the Thought Assessors as "likely to lead to eating". Your Steering Subsystem sees that and says "Boo, given my current nausea, that seems like a bad thought." It will feel a bit aversive. "Yuck, I'm *really* ordering this huge cake??" you say to yourself.

Logically, you know that *come next week*, when you actually receive the cake, you won't feel nauseous anymore, and you'll be delighted to have the cake. But still, right now, you feel kinda gross and unmotivated to order it.

Do you order the cake anyway? Sure! Maybe the value function (a.k.a. the "will lead to reward" Thought Assessor) is strong enough to overrule the effects of the "will lead to eating" Thought Assessor. Or maybe you call up a different motivation: you imagine yourself as the kind of person who has good foresight and makes good sensible decisions, and who isn't stuck in the moment. That's a *different* thought in your head, which consequently activates a *different* set of Thought Assessors, and maybe *that* gets high value from the Steering Subsystem. Either way, you do in fact call the bakery to place the cake order for next week, despite feeling nauseous right now. What a heroic act!

1. ^

   Side note: I happen to think there's something akin to "less discounting" (discount factor closer to 1.0) for the value function compared to the various other Thought Assessors, such that complicated indirect distant-in-time plans are *predominantly* driven by the value function. This guess comes from the "incentive learning" psychological literature, but that's a story for a different blog post. Anyway, it's not all-or-nothing; I figure the other assessors are at least *somewhat* relevant, even for distant plans, as in the example here.

[Intro to brain-like-AGI safety] 8. Takeaways from neuro 1/2: On AGI development

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

Part of the "Intro to brain-like-AGI safety" post series .

8.1 Post summary / Table of contents

Thus far in the series, Post #1 set up my big picture motivation: what is "brain-like AGI safety" and why do we care? The subsequent six posts (#2–#7) delved into neuroscience. Of those, Posts #2–#3 presented a way of dividing the brain into a "Learning Subsystem" and a "Steering Subsystem", differentiated by whether they have a property I call "learning from scratch". Then Posts #4–#7 presented a big picture of how I think motivation and goals work in the brain, which winds up looking kinda like a weird variant on actor-critic model-based reinforcement learning.

Having established that neuroscience background, now we can finally switch in earnest to thinking more explicitly about brain-like AGI. As a starting point to keep in mind, here's a diagram from Post #6, edited to describe brain-like AGI instead of actual brains:

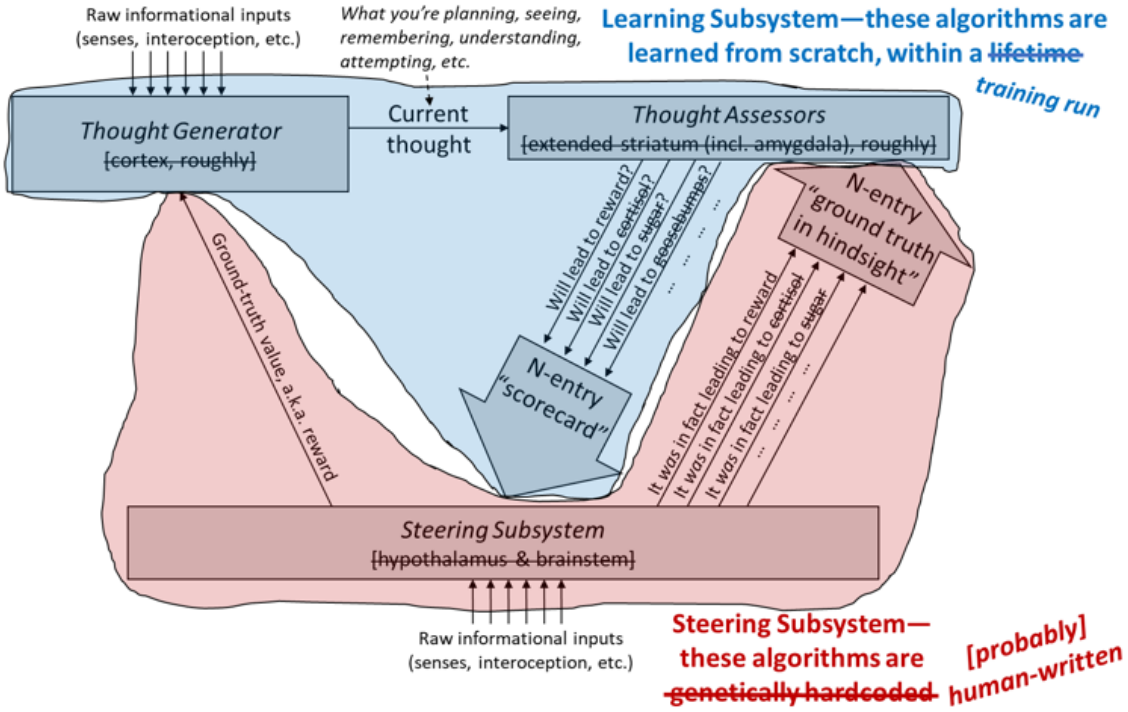


Diagram is from Post #6, with four changes to make it about brain-like-AGI rather than actual brains: (1) "lifetime" is replaced by "training run" in the top right (Section 8.2 below); (2) "genetically-hardcoded" is replaced by "[probably] human-written" in the bottom-right (Section 8.3–8.4 below); (3) references to specific brain regions like "amygdala" have been crossed out, to be replaced with bits of source code and/or sets o

trained model parameters; (4) other biology-specific words like “sugar” are crossed out, to be replaced with anything we want, as I’ll discuss in later posts.

This and the next post will extract some lessons about brain-like AGI from the discussion thus far. This post will focus on how such an AGI might be developed, and the next post will discuss AGI motivations and goals. After that, Post #10 will discuss the famous “alignment problem” (finally!), and then there will be some posts on possible paths towards a solution. Finally, in Post #15 I’ll wrap up the series with open questions, avenues for future research, and how to get involved in the field.

Back to this post. The topic is: given the discussion of neuroscience in the previous posts, how should we think about the software development process for brain-like AGI? In particular, what will be the roles of human-written source code, versus adjustable parameters (“weights”) discovered by learning algorithms?

Table of contents:

- Section 8.2 suggests that, in a brain-like AGI development process, “an animal’s lifetime” would be closely analogous to “a machine learning training run”. I discuss how long such training runs might take: notwithstanding the example of humans, who take years-to-decades to reach high levels of competence and intelligence, I claim that a brain-like AGI could plausibly have a training time as short as weeks-to-months. I also argue that brain-like AGI, like brains, will work by online learning rather than train-then-deploy, and I discuss some implications for economics and safety.
- Section 8.3 discusses the possibility of “outer-loop” automated searches analogous to evolution. I’ll argue that these are likely to play at most a minor role, perhaps for optimizing hyperparameter settings and so on, and *not* to play a major role wherein the outer-loop search is the “lead designer” that builds an algorithm from scratch, notwithstanding the fact that evolution *did* in fact build brains from scratch historically. I’ll discuss some implications for AGI safety.
- Section 8.4: While I expect the “Steering Subsystem” of a future AGI to primarily consist of human-written source code, there are some possible exceptions, and here I go through three of them: (1) There could be pre-trained image classifiers or other such modules, (2) there could be AGIs that “steer” other AGIs, and (3) there could be human feedback.

8.2 “One Lifetime” turns into “One training run”

The brain-like-AGI equivalent of “an animal’s lifetime” is “a training run”. Think of this as akin to the model training runs done by ML practitioners today.

8.2.1 How long does it take to train a model?

How long will the “training run” be for brain-like AGI?

As a point of comparison, in the human case, my humble opinion is that humans *really hit their stride* at age 37 years, 4 months, and 14 days. Everyone younger than that is a naïve baby, and everyone older than that is an inflexible old fogey. Oops, did I say “14 days”? I should have said “21 days”. You’ll have to forgive me for that error; I wrote that sentence last week, back when I was a naïve baby.

Well, whatever the number is for humans, we can ask: Will it be similar for brain-like AGIs? Not necessarily! See my post Brain-inspired AGI and the “lifetime anchor” (Sec. 6.2) for my argument that the *wall-clock* time required to train a brain-like AGI from scratch to a

powerful general intelligence is very hard to anticipate, but could plausibly wind up being as short as weeks-to-months, rather than years-to-decades.

8.2.2 Online learning implies no fundamental training-versus-deployment distinction

The brain works by online learning: instead of having multiple "episodes" interspersed by "updates" (the more popular approach in ML today), the brain is continually learning as it goes through life. I think online learning is absolutely central to how the brain works, and that any system worthy of the name "brain-like AGI" will be an online learning algorithm.

To illustrate the difference between online and offline learning, consider these two scenarios:

1. *During training*, the AGI comes across two contradictory expectations (e.g. "demand curves usually slope down" & "many studies find that minimum wage does not cause unemployment"). The AGI updates its internal models to a more nuanced and sophisticated understanding that can reconcile those two things. Going forward, it can build on that new knowledge.
2. *During deployment*, the exact same thing happens, with the exact same result.

In the online-learning, brain-like-AGI case, there's no distinction. Both of these are the same algorithm doing the same thing.

By contrast, in offline-learning ML systems (e.g. GPT-3), these two cases would be handled by two different algorithmic processes. Case #1 would involve changing the model weights, while Case #2 would not. Instead, Case #2 would solely involve changing the model *activations*.

To me, this is a huge point in favor of the plausibility of the online learning approach. It only requires solving the problem once, rather than solving it twice in two different ways. And this isn't just *any* problem; it's sorta the core problem of AGI!

I really want to reiterate what a central role online learning plays in brains (and brain-like AGIs). *A human without online learning is a human with complete anterograde amnesia.* If you introduce yourself to me as "Fred", and then 60 seconds later I refer to you as "Fred", then I can thank online learning for putting that bit of knowledge into my brain.

8.2.3 …Nevertheless, the conventional ML wisdom that "training is more expensive than deployment" still more-or-less applies

In current ML, it's common knowledge that *training is far more expensive than deployment*. For example, OpenAI allegedly spent around $10 million to train GPT-3—i.e., to get the magical list of 175 billion numbers that comprise GPT-3's weights. But now that they have that list of 175 billion numbers in hand, *running* GPT-3 is dirt cheap—last I checked, OpenAI was charging around $0.02 per page of generated text.

Thanks to online learning, brain-like AGI would have no fundamental distinction between training and deployment, as discussed in the previous section. However, the economics wind up being similar.

Imagine spending decades raising a child from birth until they were a skilled and knowledgeable adult, perhaps with advanced training in math, science, engineering, programming, etc.

Then imagine you have a sci-fi duplication machine that could instantly create 1000 copies of that adult. You send them to do 1000 different jobs. Granted, each of the copies would probably need additional on-the-job training to get up to speed. But they wouldn't need *decades* of additional training, the way it took decades of training to get them from birth to adulthood. (More discussion at Holden Karnofsky's blog.)

So, just like normal ML, there is a big fixed cost to training, and this cost can in principle be amortized over multiple copies.

8.2.4 Online learning is bad for safety, but essential for capabilities

I claim that online learning creates nasty problems for AGI safety. Unfortunately, I also claim that if we're going to build AGI at all, we need online learning, or something with similar effects. Let me elaborate on both these claims:

Online learning is bad for safety:

Let's switch to humans. Suppose I'm just now being sworn in as president of a country, and I want to always keep my people's best interests at heart, and not get drawn in by the siren song of corruption. What can I do right now, in order to control how my future self will behave? It's not straightforward, right? Maybe it's not even possible!

There just isn't a natural and airtight way for current-me to dictate what future-me will want to do. The best I can do is lots of little hacks, where I anticipate particular problems and try to preempt them. I can tie my own hands by giving an honest accountant all my bank account passwords, and asking her to turn me in if she sees anything fishy. I can have regular meetings with a trustworthy and grounded friend. Things like that may help on the margin, but again, there's no reliable solution.

In an analogous way, we can have an AGI that is *right now* trying in good faith to act ethically and helpfully. Then we keep it running for a while. It keeps thinking new thoughts, it keeps having new ideas, it keeps reading new books, and it keeps experiencing new experiences. Will it *still* be trying in good faith to act ethically and helpfully six months later? Maybe! Hopefully! But how can we be sure? This is one of many open questions in AGI safety.

(Maybe you're thinking: We could periodically boot up a snapshot of AGI-now, and give it veto-power over aspects of AGI-later? I think that's a reasonable idea, *maybe* even a good idea. But it's not a panacea either. What if AGI-later figures out how to trick or manipulate AGI-now? Or what if AGI-later has changed for the better, and AGI-now winds up holding it back? I mean, *my* younger self was a naïve baby!)

Online learning (or something with similar safety issues) is essential for capabilities:

I expect AGIs to use online learning because I think it's an effective method of making AGI—see the "solving the problem twice" discussion above (Section 8.2.2).

That said, I can imagine other possible setups that are not "online learning" *per se*, but which have similar effects, and which pose essentially the same challenges for safety, i.e. making it difficult to ensure that an initially-safe AGI continues to be safe.

I have a much harder time imagining any way to avoid those safety issues altogether. Consider:

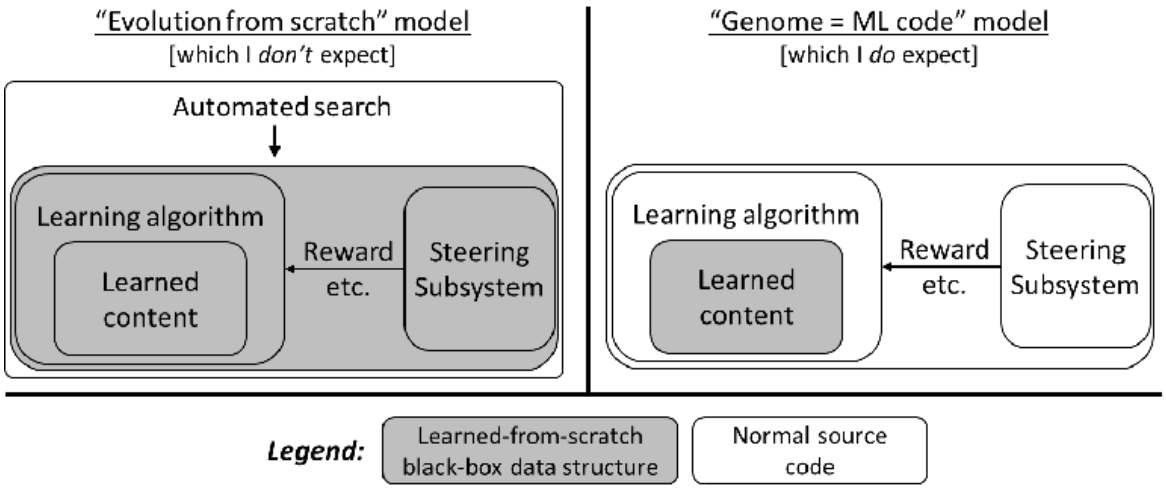
- If the AGI can think new thoughts and have new ideas and learn new knowledge "in deployment", then we would seem to be facing this goal-instability problem I'm talking

about. (See, for example, the problem of "ontological crises"; more on this in future posts.)
- If the AGI *can't* do any of those things, then is it really an AGI? Will it really be capable of doing the things we want AGI to do, like coming up with new concepts and inventing new technology? I suspect not.

8.3 Evolution-like outer-loop automated searches: maybe involved, but not the "lead designer"

"Outer loop" is a programming term for the outer of two nested control-flow loops. Here, the "inner loop" might be code that simulates a virtual animal's life, second by second, from birth to death. Then an "outer-loop search" would involve simulating lots of different animals, each with a different brain setup, in search of one that (in adulthood) displays maximum intelligence. Within-lifetime learning happens in the inner loop, whereas an outer-loop search would be analogous to evolution.

There's an extreme version of outer-loop-centric design, where (one might suppose) humans will write code that runs an evolution-like outer-loop algorithm, and this algorithm will build an AGI *from scratch*.



Two models for AGI development. The one on the left is directly analogous to how evoluti created human brains. The one on the right involves an analogy between the genome an the source code defining an ML algorithm, as spelled out in the next subsection.

The evolution-from-scratch approach (left) is discussed with some regularity in the technical AGI safety literature—see Risks From Learned Optimization and dozens of other posts about so-called "mesa-optimizers".

However, as noted in the diagram, this evolution-from-scratch approach is *not* how I expect people to build AGI, for reasons explained shortly.

That said, I'm not totally opposed to the idea of outer-loop searches; I expect them to be present with a more constrained role. In particular, when future programmers write a brain-like AGI algorithm, the source code will have a number of adjustable parameters for which it won't be obvious *a priori* what settings are optimal. These might include, for example, learning algorithm hyperparameters (such as learning rates), various aspects of neural architecture, and coefficients adjusting the relative strengths of various innate drives.

I think it's quite plausible that future AGI programmers will use an automated outer-loop search to set many or all of these adjustable parameters.

(Or not! For example, as I understand it, the initial GPT-3 training run was so expensive that it was only done once, with no hyperparameter tuning. Instead, the hyperparameters were all studied systematically in smaller models, and the researchers found trends that allowed them to extrapolate to the full model size.)

(None of this is meant to imply that learning-from-scratch algorithms don't matter for brain-like AGI. Quite the contrary, they will play a *huge* role! But that huge role will be in the *inner* loop—i.e., within-lifetime learning. See Post #2.)

8.3.1 The "Genome = ML code" analogy

In the above diagram, I used the term "genome = ML code". That refers to an analogy between brain-like AGI and modern machine learning, as spelled out in this table:

"Genome = ML code" analogy	
Human intelligence	Today's machine learning systems
Human genome	GitHub repository with all the PyTorch code for training and running the Pac-Man-playing agent
Within-lifetime learning	Training the Pac-Man-playing agent
How an adult human thinks and acts	Trained Pac-Man-playing agent
Evolution	*Maybe* the ML researchers did an outer-loop search for a handful of human-legible adjustable parameters—e.g., automated hyperparameter tuning, or neural architecture search.

8.3.2 Why I think "evolution from scratch" is less likely (as an AGI development method) than "genome = ML code"

(See also my post from March 2021: Against evolution as an analogy for how humans will create AGI .)

I think the best argument against the evolution-from-scratch model is *continuity*: "genome = ML code" is how machine learning works today. Open a random reinforcement learning paper and look at the learning algorithm. You'll see that it is human-legible, and primarily or entirely human-designed—perhaps involving things like gradient descent, TD-learning, and so on. Ditto for the inference algorithm, the reward function, etc. At most, the learning algorithm source code will have a few dozens or hundreds of bits of information that came from outer-loop search, such as the particular values of some hyperparameters, comprising a tiny share of the "design work" that went into the learning algorithm.[1]

Also, if extreme outer-loop search were really the future, I would expect that we would see today that the ML projects that rely *most* heavily on outer-loop search would be overrepresented among the most impressive, headline-grabbing, transformative results. That doesn't seem to be the case at all, as far as I can tell.

I'm merely suggesting that this pattern will continue—and for the same reason it's true today: humans are pretty good at designing learning algorithms, and meanwhile, it's extraordinarily slow and expensive to do outer-loop searches over learning algorithms.

(Granted, things that are "extraordinarily slow and expensive" today will be less so in the future. However, as time passes and future ML researchers can afford more compute, I expect that they, like researchers today, will typically "spend" that windfall on bigger models, better training procedures, and so on, rather than "spending" it on a larger outer-loop search space.)

Given all that, why do some people put a lot of stock in the "evolution-from-scratch" model? I think it comes down to the question: *Just how hard* would it be to write the source code involved in the "genome = ML code" model?

If your answer is "it's impossible", or "it would take hundreds of years", then evolution-from-scratch wins by default! On this view, even if the outer-loop search takes trillions of dollars and decades of wall-clock time and gigawatts of electricity, well, that's still the shortest path to AGI, and sooner or later some government or company will cough up the money and spend the time to make it happen.[2]

However, I *don't* think that writing the source code of the "genome = ML code" model is a hundreds-of-years endeavor. Quite the contrary, I think it's very doable, and that researchers in neuroscience & AI are making healthy progress in that direction, and that they may well succeed in the coming decades. For an explanation of why I think that, see my "timelines to brain-like AGI" discussion earlier in this series—Sections 2.8, 3.7, and 3.8.

8.3.3 Why "evolution from scratch" is worse than "genome = ML code" (from a safety perspective)

This is one of those rare cases where "what I expect to happen by default" is the same as "what I hope will happen"! Indeed, the "genome = ML code" model that I'm assuming in this series seems much more promising for AGI safety than the "evolution from scratch" model. Two reasons.

The first reason is human-legibility. In the "genome = ML code" model, the human-legibility is *bad*. But in the "evolution from scratch" model, the human-legibility is *even worse*!

In the former, the world-model is a big learned-from-scratch black-box data structure, as is the value function, etc., and we'll have our work cut out understanding their contents. In the latter, there's just one, even bigger, black box. We'll be lucky if we can even *find* the world-model, value function, and so on, *let alone* understand their contents!

The second reason, as elaborated in later posts, is that careful design of the Steering Subsystem is one of our most powerful levers for controlling the goals and motivations of a brain-like AGI, such that we wind up with safe and beneficial behavior. If we write the Steering Subsystem code ourselves, we get complete control over how the Steering Subsystem works, and visibility into what it's doing as it runs. Whereas if we use the evolution-from-scratch model, we'll have dramatically less control and understanding.

To be clear, AGI safety is an unsolved problem even in the "genome = ML code" case. I'm saying that the evolution-from-scratch AGI development approach would seemingly make it even worse.

(*Note for clarity:* this discussion is assuming that we wind up with "brain-like AGI" in either case. I'm not making any claims about brain-like AGI being more or less safe than non-brain-like AGI, assuming the latter exists.)

8.3.3.1 Is it a good idea to build human-like social instincts by evolving agents in a social environment?

A possible objection I sometimes hear is something like: "Humans aren't so bad, and evolution designed *our* Steering Subsystems, right? Maybe if we do an evolution-like outer-loop search process in an environment where multiple AGIs need to cooperate, they'll wind up with altruism and other such nice social instincts!" (I think this kind of intuition is the motivation behind projects like DeepMind Melting Pot.)

I have three responses to that.

- First, my impression (mainly from reading Richard Wrangham's *The Goodness Paradox*) is that there are huge differences between human social instincts, and chimpanzee social instincts, and bonobo social instincts, and wolf social instincts, and so on. For example, chimpanzees and wolves have dramatically higher "reactive aggression" than humans and bonobos, though all four are intensely social. The evolutionary pressures driving social instincts are a sensitive function of the power dynamics and other aspects of social groups, possibly with multiple stable equilibria, in a way that seems like it would be hard to control by tweaking the knobs in a virtual environment.
- Second, if we set up a virtual environment where AGIs are incentivized to cooperate with AGIs, we'll get AGIs that have cooperative social instincts *towards other AGIs in their virtual environment*. But what we *want* is AGIs that have cooperative social instincts *towards humans in the real world*. A Steering Subsystem that builds the former might or might not build it in a way that generalizes to the latter. Humans, I note, are often compassionate toward their friends, but rarely compassionate towards members of an enemy tribe, or towards factory-farmed animals, or towards large hairy spiders.
- Third, human social instincts leave something to be desired! For example, it has been argued (plausibly in my opinion) that a low but nonzero prevalence of psychopathy in humans is not a random fluke, but rather an advantageous strategy from the perspective of selfish genes as studied by evolutionary game theory. Likewise, evolution seems to have designed humans to have jealousy, spite, teenage rebellion, bloodlust, and so on. And *that's* how we want to design our AGIs?? Yikes.

8.4 Other non-hand-coded things that might go in a future brain-like-AGI Steering Subsystem

As discussed in Post #3, I claim that the Steering Subsystem in mammal brains (i.e., hypothalamus and brainstem) consists of genetically-hardcoded algorithms. (For discussion and caveats, see Post #2, Section 2.3.3.)

When we switch to AGI, my corresponding expectation is that future AGIs' Steering Subsystems will consist of primarily human-written code—just as today's RL agents typically have human-written reward functions.

However, it may not be *completely* human-written. For one thing, as discussed in the previous section, there may be a handful of adjustable parameters set by outer-loop search, e.g. coefficients controlling the relative strengths of different innate drives. Here are three

other possible exceptions to my general expectation that AGI Steering Subsystems will consist of human-written code.

8.4.1 Pre-trained image classifiers, etc.

Plausibly, an ingredient in AGI Steering Subsystem code could be something like a trained ConvNet image classifier. This would be analogous to how the human superior colliculus has something-like-an-image-classifier for recognizing a prescribed set of innately-significant categories, like snakes and spiders and faces (see Post #3, Section 3.2.1). Likewise, there could be trained classifiers for audio or other sensory modalities.

8.4.2 A tower of AGIs steering AGIs?

In principle, in place of the normal Steering Subsystem, we could have a *whole separate AGI* that is watching the thoughts of the Learning Subsystem and sending appropriate rewards.

Heck, we could have a whole *tower* of AGIs-steering-AGIs! Presumably the AGIs would get more and more complex and powerful going up the tower, gradually enough that each AGI is up to the task of steering the one above it. (It could also be a pyramid rather than a tower, with multiple dumber AGIs collaborating to comprise the Steering Subsystem of a smarter AGI.)

I don't think this approach is necessarily useless. But it seems to me that I still haven't even gotten past the first step, where we make *any* safe AGI. Building a tower of AGIs-steering-AGIs does not avert the need to make *a* safe AGI in a different way. After all, the tower needs a base!

Once we solve that first big problem, *then* we can think about whether to use that new AGI directly to solve human problems, or to use it indirectly, by having it steer even-more-powerful AGIs, analogously to how we humans are trying to steer the first AGI.

Of those two possibilities, I lean towards "use that first AGI directly" being a more promising research direction than "use that first AGI to steer a second, more powerful, AGI". But I could be wrong. Anyway, we can cross that bridge when we get to it.

8.4.3 Humans steering AGIs?

If an AGI's Steering Subsystem can (maybe) be another AGI, then why can't it be a human?

Answer: if the AGI is running at human brain speed, maybe it would be thinking 3 thoughts per second (or something). Each "thought" would need a corresponding reward and maybe dozens of other ground-truth signals. A human would never be able to keep up!

What we *can* do is have human feedback be an *input* into the Steering Subsystem. For example, we could give the humans a big red button that says "REWARD". (We probably *shouldn't*, but we *could*.) We can also have other forms of human involvement, including ones with no biological analog—we should keep an open mind.

1. ^

   For example, here's a random neural architecture search (NAS) paper: "The evolved transformer". The authors brag about their "large search space", and it *is* a large search space *by the standards of NAS*. But searching through that space still yields only 385 bits of information, and the end result fits in one easily-human-legible

diagram in the paper. By contrast, the weights of an ML trained model may easily comprise millions or billions of bits of information, and the end result requires heroic effort to understand. We can also compare those 385 bits to the number of bits of information in the *human-created* parts of the learning algorithm source code, such as the code for matrix multiplication, softmax, autograd, shuttling data between the GPU and the CPU, and so on. The latter parts comprise orders of magnitude more than 385 bits of information. This is what I mean when I say that things like hyperparameter tuning and NAS contribute a tiny proportion of the total "design work" in a learning algorithm.

(The most outer-loop-search-reliant paper that I know of is AutoML-Zero, and even there, the outer-loop search contributed effectively 16 lines of code, which the authors had no trouble understanding.)

2. ^

If you're curious for some ballpark estimates of how much time and money would it take to perform an amount of computation equivalent to the entire history of animal evolution on Earth, see the "Evolution anchor" discussion in Ajeya Cotra's 2020 draft report on biological anchors. Obviously, this is not exactly the same as the amount of computation required for evolution-from-scratch AGI development, but it's not *entirely* irrelevant either. I won't talk about this topic more; I don't think it's important, because I don't think evolution-from-scratch AGI development will happen anyway.

[Intro to brain-like-AGI safety] 9. Takeaways from neuro 2/2: On AGI motivation

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

Part of the "Intro to brain-like-AGI safety" post series .

9.1 Post summary / Table of contents

Most posts in the series thus far—Posts #2–#7—have been primarily about neuroscience. Then, starting with the previous post, we've been applying those ideas to better understand brain-like-AGI safety (as defined in Post #1).

In this post, I'll discuss some topics related to the motivations and goals of a brain-like AGI. Motivation is of paramount importance for AGI safety. After all, our prospects are a heck of a lot better if future AGIs are *motivated* to bring about a wonderful future rich in human flourishing, compared to if they're *motivated* to kill everyone. To get the former and not the latter, we need to understand how brain-like-AGI motivation works, and in particular how to point it in one direction rather than another. This post will cover assorted topics in that area.

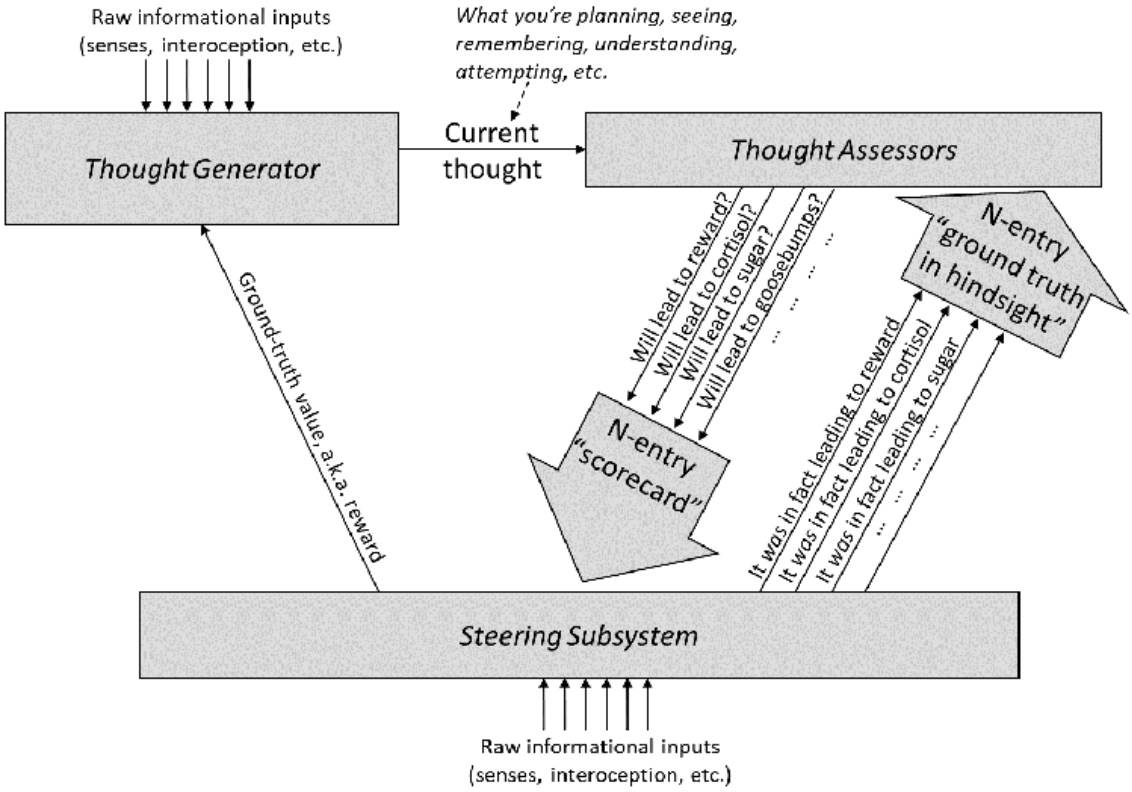
Table of contents:

- Section 9.2 argues that the goals and preferences of a brain-like AGI are defined in terms of latent variables in its world-model. These can be *related* to outcomes, actions, or plans, but are not exactly any of those things. Also, the algorithms generally don't distinguish between instrumental and final goals.
- Section 9.3 has a deeper discussion of "credit assignment", which I previously introduced with an example back in Post #7 (Section 7.4). "Credit assignment", as I use the term in this series, is a synonym of "updates to the Thought Assessors", and is the process whereby a concept (= latent variable in the world-model) can get "painted" with positive or negative valence, and/or start triggering involuntary visceral reactions (in the human case). This type of "credit assignment" is a key ingredient in how an AGI could wind up wanting to do something.
- Section 9.4 defines "wireheading". An example of "wireheading" would be if an AGI hacks into itself and sets the "reward" register in RAM to the maximum possible value. I will argue that brain-like AGI will "by default" have a "weak wireheading drive" (desire to wirehead, other things equal), but probably not a "strong wireheading drive" (viewing wireheading as the best possible thing to do, and worth doing at any cost).
- Section 9.5 spells out an implication of the wireheading discussion above: brain-like AGI is generally *NOT* trying to maximize its future reward. I give a human example, and then relate it to the concept of "observation-utility agents" in the literature.
- Section 9.6 argues that, in brain-like AGI, the Thought Assessors mediate a relationship between motivation and neural network interpretability. For example, the assessment "This thought/plan is likely to lead to eating" is simultaneously (1) a data-point contributing to the interpretability of the thought/plan within the learned world-model, and (2) a signal that we should carry out that thought/plan if we're hungry. (This point applies to any reinforcement learning system compatible with multi-dimensional value functions, not just "brain-like" ones. Ditto for the next bullet point.)
- Section 9.7 describes how we might be able to "steer" the AGI's motivations in real-time, and how this steering would impact not just the AGI's immediate actions but also its long-term plans and "deep desires".

9.2 The AGI's goals and desires are defined in terms of latent variables (learned concepts) in its world-model

Do you like football? Well, "football" is a learned concept living inside your world-model. Learned concepts like that are the only kinds of things that it's *possible* to "like". You cannot like or dislike [nameless pattern in sensory input that you've never conceived of]. It's possible that you would find this nameless pattern *rewarding*, were you to come across it. But you can't *like* it, because it's not currently part of your world-model. That also means: you can't and won't make a goal-oriented plan to induce that nameless pattern.

I think this is clear from introspection, and I think it's equally clear in our motivation picture (see Posts #6–#7). There, I used the term "thought" in a broad sense to include everything in conscious awareness and more—what you're planning, seeing, remembering, understanding, attempting, etc. A "thought" is what the Thought Assessors assess, and it is built out of some configuration of the learned latent variables in your generative world-model.



Our motivation model—see Post #6 for discussion

Why is it important that an AGI's goals are defined in terms of latent variables in its world-model? Lots of reasons! It will come up over and over in this and future posts.

9.2.1 Implications for "value alignment" with humans

The above observation is one reason that "value alignment" between a human and an AGI is an awful mess of a problem. A brain-like AGI will have latent variables in its learned world-model, while a human has latent variables in *their* learned world-model, but they are different world-models, and the latent variables in one may have a complex and problematic relationship to the latent variables in the other. For example, the human's latent variables could include things like "ghosts" that don't really correspond to *anything* in the real world! For more on this topic, see John Wentworth's post The Pointers Problem.

(I won't say much about "defining human values" in this series—I want to stick to the narrower problem of "avoiding catastrophic AGI accidents like human extinction", and I don't think a deep dive into "defining human values" is necessary for that. But "defining human values" would still a good thing to do, and I'm happy for people to be working on it—see for example 1,2.)

9.2.2 Preferences are over "thoughts", which can *relate to* outcomes, actions, plans, etc., but are different from all those things

Thought Assessors assess and compare "thoughts", i.e. configurations of an agent's generative world-model. The world-model is imperfect—a complete understanding of the world is far too complex to fit in any brain or silicon chip. Thus a "thought" inevitably involves attending to some things and ignoring others, conceptualizing things in certain ways, matching things to the nearest-available category even if it's not a perfect fit, etc.

Some implications:

- You can conceptualize a single sequence of motor actions in many different ways, and it will be more or less appealing depending on how you're thinking about it: consider the thought "I'm gonna go to the gym" versus the thought "I'm gonna go to the gym *to get ripped*". See (Brainstem, Neocortex) ≠ (Base Motivations, Honorable Motivations) for related discussion.
- Similarly, you can conceptualize a single future state of the world in many different ways, e.g. by attending to different aspects of it, and it will thereby become more or less appealing. This can lead to circular preferences; I put an example in this footnote[1].
- A thought can concern immediate actions, and future actions, and semantic context, and expectations of what will happen while we're doing the thing, and expectations of what will result after we finish doing the thing, etc. Thus we can have "consequentialist" preferences about future states, or "deontological" preferences about actions, etc. For example, the thought "I'm going to go to the store, and then I'll have milk" includes action-related "I'm going to go to the store" neurons, *and* consequence-related "I'll have milk" neurons; the Thought Assessors and Steering Subsystem can endorse or reject the thought based on *either* of those. See Consequentialism & Corrigibility for more on this topic.
- None of this is meant to imply that a brain-like AGI can't approximate an ideal rational consequentialist utility-maximizer! Just that this would be a property of a particular trained model, rather than inherent in the AGI's source code. For example, a brain-like AGI can read *The Sequences* (just like a human can), and it can internalize those lessons into a set of learned metacognitive heuristics that catch and correct faulty intuitions and habits-of-thought that undermine effectiveness[2] (just like a human can), and the AGI may in fact want to actually do this for the same reasons that a human might read *The Sequences*, ~~namely to submit to a 30-hour-long hazing ritual and thus earn in-group membership~~[3] namely because it wants to think clearly and accomplish its goals.

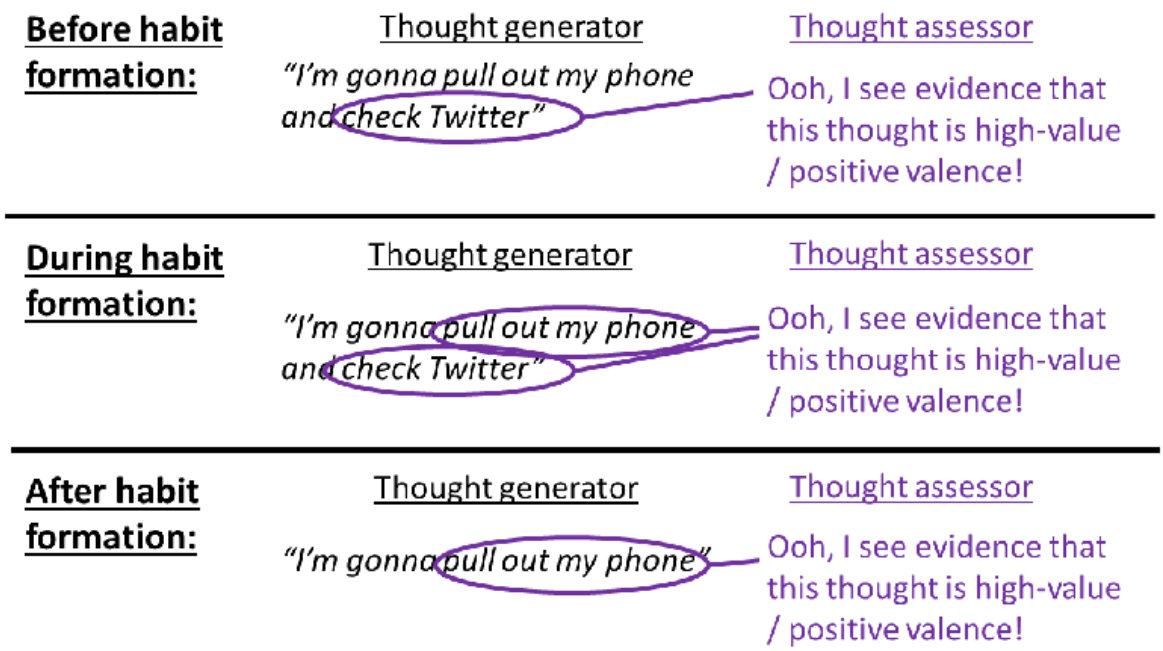
9.2.2 Instrumental & final preferences seem to be mixed together

There's an *intuitive* sense in which we have instrumental preferences (things we prefer because they have typically been useful in the past as a means to an end—e.g., I prefer wearing a watch because it helps me check the time), and final preferences (things we prefer as an end in themselves—e.g., I like feeling good, and dislike getting mauled by a bear). For example, Spencer Greenberg did a survey where some participants, but not others, described "there are beautiful things in the world" as a final goal—they cared about there being beautiful things, even if those things were located deep underground where no conscious being would ever see them. Do you agree or disagree? To me, the most interesting thing is that some people will answer: "I don't know, I've never thought about that before, hmm, give me a second." I think there's a lesson here!

Namely: It seems to me that there is *not* a distinction between instrumental and final preferences baked deeply into brain algorithms. If you think a thought, and your Steering Subsystem endorses it as a high-value thought, I think the computation looks the same if it's a high-value thought for instrumental reasons, versus a high-value thought for final reasons.

I should clarify: You can do instrumental *things* without them being an instrumental *preference*. For example, when I first got a smartphone, I would sometimes take it out of my pocket to check Twitter. At the time, I had no preference for pulling out my cell phone *per se*. Instead, I was thinking a thought along the lines of: "I'm going to pull out my cell phone and then check Twitter." The Steering Subsystem endorses *this* as a high-value thought, but only because of the *second* part of the thought, the part that involves checking Twitter.

Then after a while, "credit assignment" (next section) worked its magic and put a *new* preference into my brain, a preference for reaching into my pocket and pulling out my cell phone *per se*. After that, I started pulling out my cell phone without having any idea why. And *now* it's an "instrumental preference".



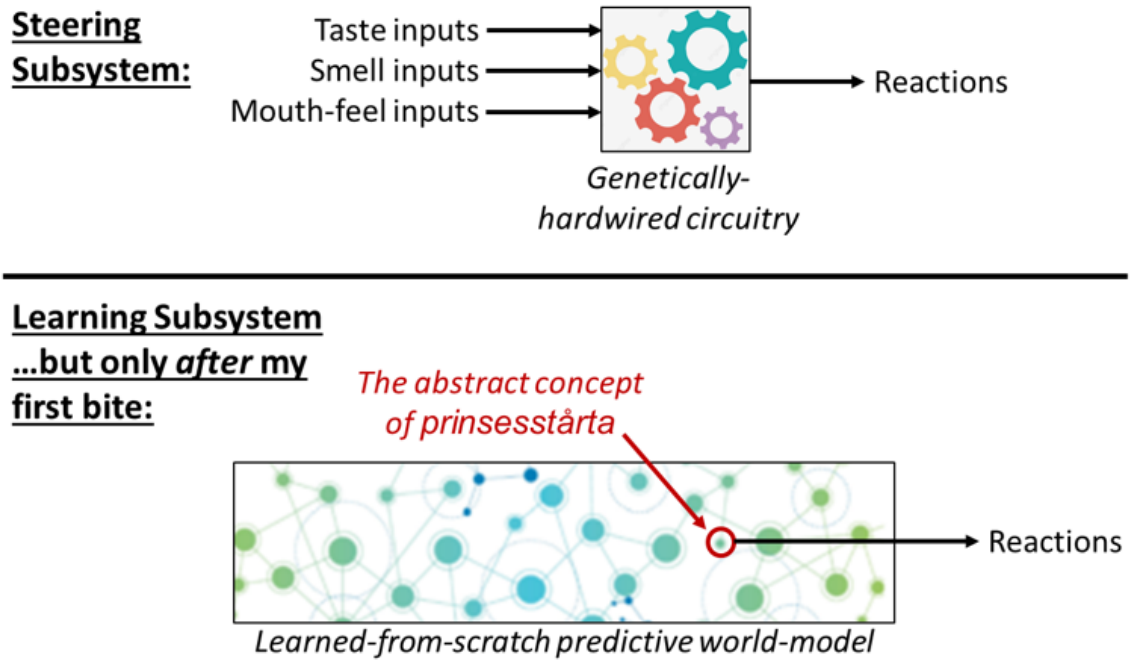
Habit-formation is the process whereby credit-assignment turns an instrumental *behavior* into an instrumental *preference*.

(Note: Just because instrumental and final preferences are mixed up in human brains doesn't mean they *have* to be mixed up in brain-like AGIs. For example, I can vaguely imagine some system for flagging positive-valence concepts with some explanation for how they came to be positive-valence. In the example above, maybe we could wind up with a dotted line from some innate drive to the "Twitter" concept, and then another dotted line from the "Twitter" concept to the "reach into my pocket and grab my phone" concept. I presume the dotted lines would probably be functionally inert for AGI operations, but it would be great to have them available to help with neural network interpretability. To be clear, I don't know if this could really work as described; I'm just brainstorming.)

9.3 "Credit assignment" is how latent variables get painted with valence

9.3.1 What is credit assignment?

I introduced the idea of "credit assignment" in Post #7 (Section 7.4), and I suggest re-reading that now, so that you have a concrete example in mind. Recall this diagram:



> Copied from Post #7, see there for context.

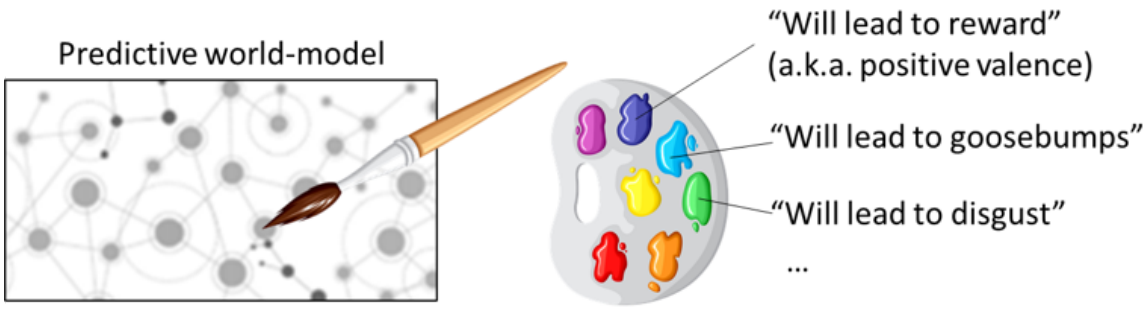
As a reminder, the brain has "Thought Assessors" (Post #5 & #6) that work by supervised learning (with the supervisory signals coming from the Steering Subsystem). Their role is to translate from latent variables (a.k.a. concepts) in the world model ("paintings", "taxes", "striving", etc.) to parameters that the Steering Subsystem can understand (arm pain, blood sugar levels, grimacing, etc.). For example, when I took a bite of cake in Post #7, a world-model concept ("myself eating prinsesstårta cake") got attached to genetically-meaningful variables (sweet taste on my tongue, reward, etc.).

I'm calling that process "credit assignment"—in the sense that the abstract concept of "myself eating prinsesstårta cake" *gets credit* for the sweet taste on my tongue.

Kaj Sotala has a kinda poetic description of what I call credit assignment here:

> Mental representations … [are] imbued with a context-sensitive affective gloss.

I find myself visualizing a fine-tip paintbrush painting positive valence onto my mental concept of prinsesstårta. Besides the "valence" paint, there are various other paint colors associated with other visceral reactions.



I sometimes like to visualize credit assignment as kinda like "painting" the latent variables in your predictive world-model with associations to rewards and other innate reactions.

Credit assignment can work in funny ways. Lisa Feldman Barrett tells a story where one time she went on a date, felt butterflies in her stomach, and thought she had found True Love—only to discover later that evening that she was coming down with the flu! Likewise, if I'm pleasantly surprised to win a prize, my brain can "assign credit" to my hard work and skill, or it can "assign credit" to the fact that I'm wearing my lucky underwear.

I said "my brain can assign credit" instead of "I can assign credit" just now, because I don't want to imply that this is a voluntary choice that I made. Instead, credit assignment is some dumb algorithm in the brain. Speaking of which:

9.3.2 How does credit assignment work?—the short answer

If credit assignment is a dumb algorithm in the brain, *exactly what* dumb algorithm is it?

I think, at least to a first approximation, it's the obvious one:

Whatever thought is active right now gets the credit.

That's "obvious" in the sense that the Thought Assessors are using supervised learning (see Post #4), and this is what supervised learning would do by default. After all, the "context" inputs to the Thought Assessors are describing whatever thought is active right now, so if we do a gradient-descent update on the error (or something functionally similar to a gradient-descent update), this "obvious" algorithm is what we'll get.

9.3.3 How does credit assignment work?—fine print

I think it's worth investing a bit more time on this topic, because credit assignment is central to AGI safety—after all, it's how a brain-like AGI would wind up wanting some things rather than others. So I'll just list out some assorted thoughts about how it works in humans.

1. Credit assignment can have "priors" that bias what type of concept gets what type of credit:

Recall from Posts #4–#5 that each Thought Assessor has its own "context" signals that serve as inputs to its predictive model. Imagine that some specific Thought Assessor has *only* context data from the visual cortex, for example. It will be forced to "assign credit" to the primarily-visual patterns stored in that part of the neural architecture—as if it had a 100%-confident "prior" that *only* the visual cortex's stored patterns could *possibly* be helpful for the prediction task.

Naïvely, we might think this kind of "prior" is always a bad idea: the more different context signals that a Thought Assessor has, the better its predictive models will be, right? Why restrict them? Two reasons. First, a good prior will lead to faster learning. Second, the Thought Assessors are just one component of a larger system. We shouldn't take for granted that a more-predictively-accurate Thought Assessor is necessarily a good thing *for the larger system*.

Here's a famous example of these kinds of "priors" in psychology: rats can easily learn to freeze in response to a *sound* that precedes an electric shock, and rats can easily learn to feel nauseous in response to a *taste* that precedes a bout of vomiting. But not vice-versa! This might reflect, for example, a brain architectural design feature wherein the nausea-predicting Thought Assessor has taste-related context (e.g. from the insular cortex) but not audiovisual-related context (e.g. from the temporal lobe), and vice-versa for the freeze-predicting Thought Assessor. (More on the nausea example shortly.)

2. Credit assignment is very sensitive to timing:

Above I suggested "Whatever thought is active right now gets the credit". But I didn't say what "right now" means.

Example: Suppose I'm walking down the street, thinking about the TV show that I watched last night. Suddenly I have a sharp pain on my back—somebody punched me. Two things happen in my brain, almost immediately:

1. My thoughts and attention turn to this new pain in my back (possibly including some generative model of its causes),
2. My brain does the "credit assignment" thing, where some concepts in my world-model gets viscerally associated with this new pain sensation.

The trick is, we want (A) to happen *before* (B)—otherwise, I'll wind up with a visceral anticipation of back pain whenever I think about that TV show that I watched last night.

I do in fact think that the brain is able to ensure that (A) happens before (B), at least by and large. (I might get a *bit* of a spurious association with the TV show.)[4]

3. …And timing can interact with "priors" too!

Conditioned Taste Aversion (CTA) is a phenomenon where, if I get nauseous right now, it causes an aversion to whatever tastes I was exposed to a few hours earlier—not a few seconds earlier, not a few days earlier, just a few hours earlier. (I alluded to CTA above, but not its timing aspect.) The evolutionary reason for this is straightforward: a few hours is presumably how long it typically takes for a toxic food to induce nausea. But how does it work mechanistically?

The insular cortex is the home of neurons that form a generative model of taste sensory inputs. According to "A molecular mechanism underlying gustatory memory trace for an association in insular cortex" by Adaikkan & Rosenblum (2015), these neurons have molecular mechanisms that put them in a special flagged state for the subsequent several hours after they fire.

Then the rule I suggested above ("Whatever thought is active right now gets the credit") needs to be modified to: "Whatever neurons are in that special flagged state right now get the credit."

4. Credit assignment has a "Finders Keepers" characteristic:

Once you have a way to accurately predict some set of supervisory signals, it makes the corresponding error signal go away, so we stop assigning more credit in those situations. So I think the *first* good predictive model that our brain comes across, gets to stick around by default. I think this is related to blocking in behaviorist psychology.

5. The Thought Generator doesn't have direct voluntary control over credit assignment, but it probably has at least *some* ability to manipulate it

There's a sense in which the Thought Generator and Thought Assessors are in an adversarial relationship, i.e. working at cross-purposes. In particular, they are trained to optimize different signals.[5] For example, one time my boss yelled at me, and I very much didn't want to start crying, but my Thought Assessors assessed that it was an appropriate time to cry, and so I did![6] Given that adversarial relationship, I have a strong presumption that the Thought Generator is not set up to have direct ("voluntary") control over credit assignment. This also seems to match introspection.

On the other hand, "no direct voluntary control" is quite different from "no control at all". Again, I don't have direct voluntary control over crying, but I can nevertheless summon tears, at least a little bit, via the roundabout strategy of imagining baby kittens shivering in the cold rain (Post #6, Section 6.3.3).

So, suppose I currently hate X, but I want to *will myself* to really like X. It seems to me that this task is not straightforward, but also that it's not impossible. It may take some self-reflective skill, mindfulness, planning, and so on, but if the Thought Generator thinks just the right thoughts at the right time, it can probably pull it off.

And an AGI might have an easier time than a human! After all, unlike in humans, an AGI may be able to literally hack into its own Thought Assessor, and change the settings however it likes. And that nicely transitions us to the next topic…

9.4 Wireheading: possible but not inevitable

9.4.1 What is wireheading?

The concept of "wireheading" gets its name from the idea of sticking a wire into a certain part of your brain, and running current through it. If you do it right, it could directly elicit ecstatic pleasure, deep satisfaction, or other nice feelings, depending on the exact part of the brain that the wire is in. Wireheading can be a much easier way to elicit those nice feelings, compared to, y'know, finding True Love, cooking the perfect soufflé, winning the praise of your childhood hero, and the like.

In the classic, nightmare-inducing, wireheading experiment (see "Brain Stimulation Reward"), a wire in a rat's brain is activated when the rat presses a lever. The rat will press the lever over and over, not stopping to eat or drink or rest, even for 24 hours straight, until eventually collapsing from exhaustion. (ref)

Anyway, the concept of wireheading has been analogized to AI. The idea here is that a reinforcement learning agent is designed to maximize its reward. So, maybe it will hack into

its own RAM, and overwrite the "reward" register to infinity! Next I'll talk about whether that's likely to happen, and then how worried we should be if it does.

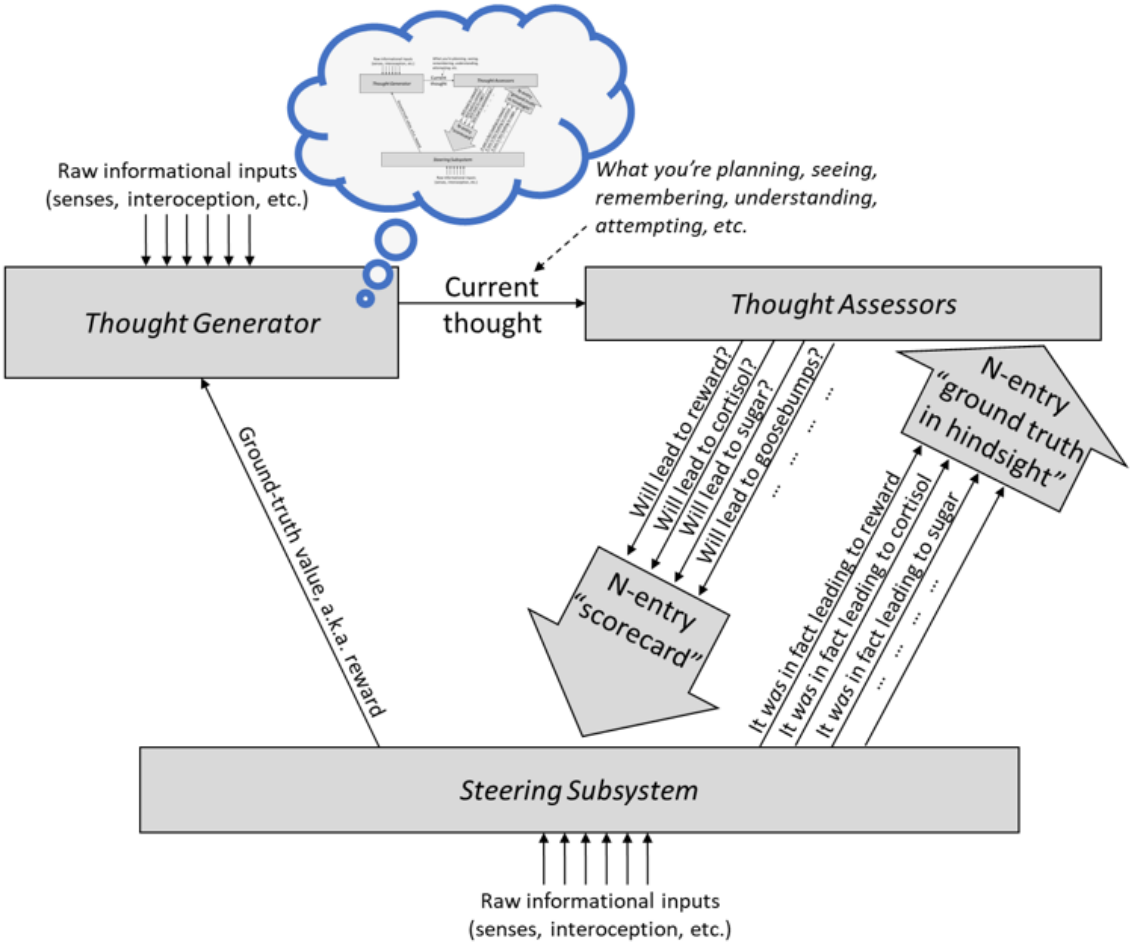
9.4.2 Will brain-like AGIs want to wirehead?

Well, first, do *humans* want to wirehead? I need to distinguish two things:

- ***Weak wireheading drive:*** "I want a higher reward signal in my brain, other things equal."
- ***Strong wireheading drive:*** "I want a higher reward signal in my brain—and I would do *anything* to get it."

In the human case, maybe we can equate a wireheading drive with "the desire to feel good", i.e. hedonism.[7] If so, it would suggest that (almost) all humans have a "weak wireheading drive" but not a "strong wireheading drive". We want to feel good, but we generally care at least a *little* bit about other things too.

How do we make sense of that? Well, think of the previous two sections above. For a human to want reward: *first*, it needs to have a reward concept in its world-model, and *second*, credit assignment needs to flag that concept as being "good". (I'm using the term "reward concept" in a broad sense that also would also include a "feeling good" concept.[7])



An AGI (or human) can have *self-reflective* concepts, and hence can be motivated to fut with its internal settings and operations.

Given that, and the notes on credit assignment in Section 9.3 above, I figure:

- *Avoiding a strong wireheading drive* is trivial and automatic; it just requires that credit assignment has, *at least once ever*, assigned positive-valence-related credit to *anything* other than the reward / feeling good concept.
- *Avoiding a weak wireheading drive* seems quite tricky. Maybe we could minimize it using timing and priors (Section 9.3.3 above), but avoiding it altogether would, I presume, require special techniques—I vaguely imagine using some kind of interpretability technique to find the reward / feeling good concept in the world-model, and manually disconnecting it from any Thought Assessors, or something like that.

(There's also a possibility that a weak-wireheader will *self-modify* into a strong-wireheader; more on that kind of thing in the next post.)

9.4.3 Wireheading AGIs would be dangerous, not merely unhelpful

There's an unhelpful intuition that trips up many people: When we imagine a wireheading AGI, we compare it to a human in the midst of an intense recreational drug high. Such a human is certainly not methodically crafting, revising, and executing a brilliant, devious plan to take over the world. While they're high, they're probably just closing their eyes and feeling good, or maybe they're dancing or something; it depends on the drug. So this intuition suggests that wireheading is a capabilities problem, but not a catastrophic accident risk.

I think there's a kernel of truth to this intuition: as discussed in Posts #6–#7, reward / value signals guide cognition and planning, so if reward gets stuck onto a very positive setting, cognition and planning become impossible.

But it's wrong to draw the conclusion that wireheading is not a catastrophic accident risk. [8] Consider what happens *before* the AGI starts wireheading. If it entertains the plan "I will wirehead", that thought would presumably get a high value from the Steering Subsystem. But if it thinks about it a bit more, it would realize that its expectation should be "I will wirehead for a while, and then the humans will shut me down and repair the memory leak so that I can't wirehead anymore." *Now* the plan doesn't sound so great! So the AGI may come up with a better plan, one that involves things like seizing control of its local environment, and/or the power grid, and/or the whole world, and/or building itself a "bodyguard AI" that does all those things for it while it wireheads, etc. So really, I think wireheading *does* carry a risk of catastrophic accidents, including even the kinds of human-extinction-level accident risks that I discussed in Post #1.

9.5 AGIs do NOT judge plans based on their expected future rewards

This directly follows from the previous section, but I want to elevate it to a top-level heading, as "AGIs will try to maximize future rewards" is a common claim.

If the Thought Generator proposes a plan, the Thought Assessors will evaluate its likely consequences according to their *current* models, and the Steering Subsystem will endorse or reject the plan largely on that basis. Those current models need not align with "expected future rewards".

The Thought Generator's predictive world-model can even "know" about some discrepancy between "expected future rewards" and the Thought Assessor's assessment of expected

future reward. It doesn't matter! The Thought Assessor's assessments won't automatically correct themselves, and will still continue to determine what plans the AGI will execute.

9.5.1 Human example

Here's a human example. I'll talk about cocaine instead of wireheading. (They're not so different, but cocaine is more familiar.)

True fact: I've never done cocaine. Suppose I think to myself right now "maybe I'll do cocaine". *Intellectually*, I'm confident that if I did cocaine, I would have, umm, lots of very intense feelings. But *viscerally*, imagining myself doing cocaine is mostly neutral! It doesn't make me feel much of anything in particular.

So for me right now, my *intellectual* expectations (of what would happen if I did cocaine) are out of sync with my *visceral* expectations. Apparently my Thought Assessors took a look at the thought "maybe I'll do cocaine", and collectively shrugged: "Nothing much going on here!" Recall that the Thought Assessors work by credit assignment (Section 9.3 above), and apparently the credit assignment algorithm just doesn't update strongly on hearsay about what cocaine feels like, nor does it update strongly on my reading neuroscience papers about how cocaine binds to dopamine transporters.

By contrast, the credit assignment algorithm *does* update strongly on a direct, first-person experience of intense feelings.

And thus, people can get addicted to cocaine after *using* cocaine, whereas people *don't* get addicted to cocaine after *reading about* cocaine.

9.5.2 Relation to "observation-utility agents"

For a more theoretical perspective, here is Abram Demski (sorry for the jargon—if you don't know what AIXI is, don't worry, you can still probably get the gist):

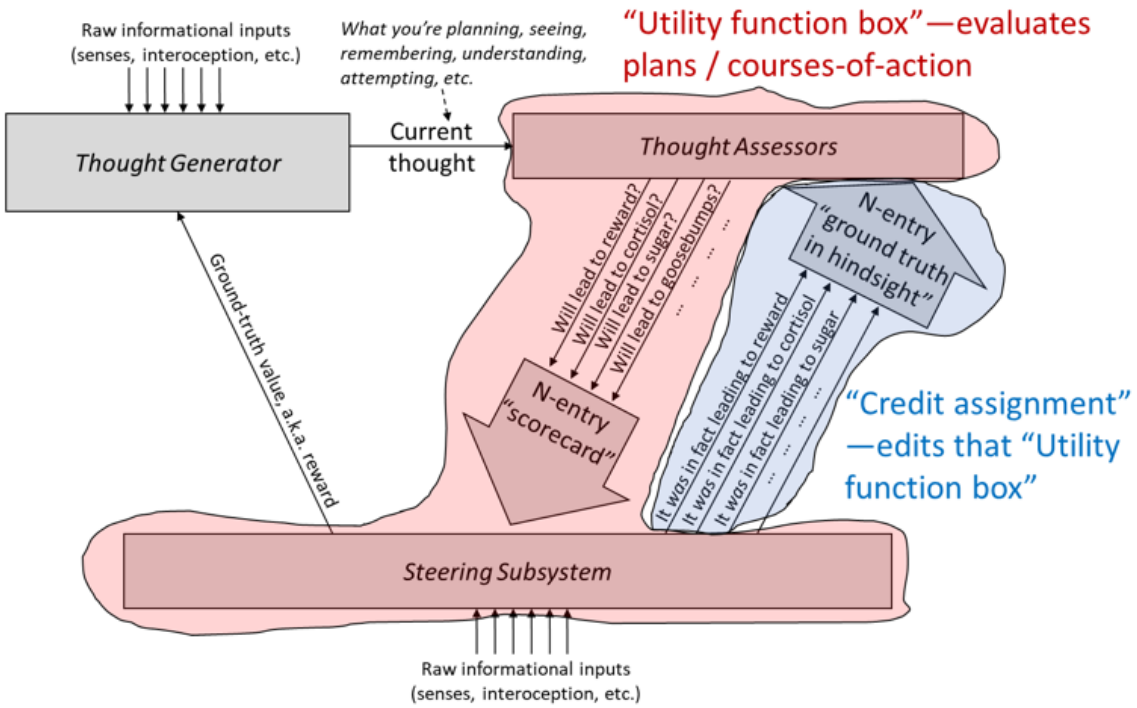
> As a first example, consider the wireheading problem for AIXI-like agents in the case of a fixed utility function which we know how to estimate from sense data. As discussed in Daniel Dewey's Learning What to Value and other places, if you try to implement this by putting the utility calculation in a box which rewards an AIXI-like RL agent, the agent can eventually learn to modify or remove the box, and happily does so if it can get more reward by doing so. This is because the RL agent predicts, and attempts to maximize, reward received. If it understands that it can modify the reward-giving box to get more reward, it will.
>
> We can fix this problem by integrating the same reward box with the agent in a better way. Rather than having the RL agent learn what the output of the box will be and plan to maximize the output of the box, we use the box *directly* to evaluate possible futures, and have the agent plan to maximize that evaluation. Now, if the agent considers modifying the box, it evaluates that future *with the current box*. The box as currently configured sees no advantage to such tampering. This is called an observation-utility maximizer (to contrast it with reinforcement learning)….
>
> This feels much like a use/mention distinction. The RL agent is maximizing "the function in the utility module", whereas the observation-utility agent (OU agent) is maximizing the function in the utility module.

Our brain-like AGI, despite being "RL",[9] is really closer to the "observation-utility agent" paradigm: the Thought Assessors and Steering Subsystem work together to evaluate plans / courses-of-action, just as Abram's "box" does.

However, the brain-like AGI has an additional twist that the Thought Assessors get gradually updated over time by "credit assignment" (Section 9.3 above).

Thus we wind up with something *vaguely* like the following:

- A utility-maximizing agent
- …plus a process that occasionally updates the utility function, in a way that tends to make it better match a reward function.



This diagram spells out how our brain-like-AGI motivation picture fits into the "observation-utility agent" paradigm, as described in the text.

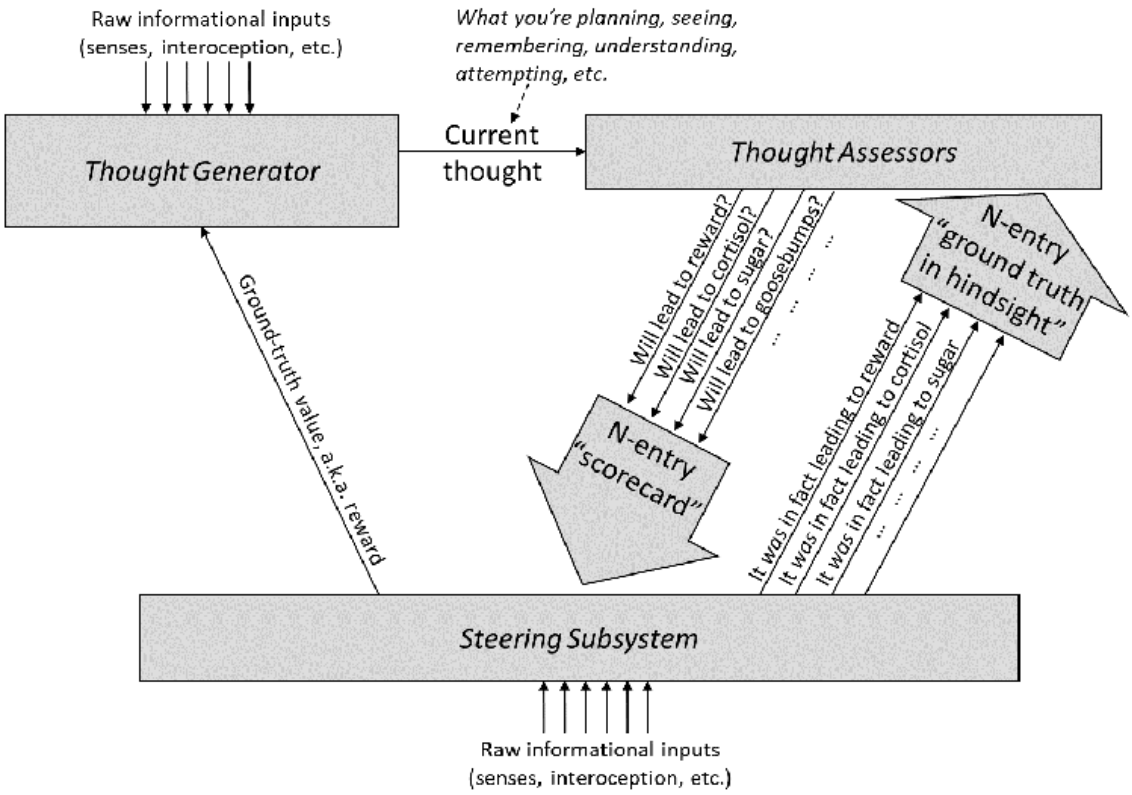
Note that we don't *want* the credit assignment process to perfectly "converge"—i.e., to reach a place where the utility function perfectly matches the reward function (or in our terminology, reach a place where the Thought Assessors never get updated because they evaluate plans in a way that always perfectly matches the Steering Subsystem).

Why don't we want perfect convergence? Because perfect convergence would lead to wireheading! And wireheading is bad and dangerous! (Section 9.4.3 above.) Yet at the same time, we need *some* amount of convergence, because the reward function is supposed to be sculpting the AGI's goals! (Remember, the Thought Assessors start out random and hence useless.) It's a Catch-22! I'll return to this topic in the next post.

(Astute readers may have also noticed another problem: the utility-maximizer may try to maintain its goals by sabotaging the credit-assignment process. I'll elaborate on that in the next post as well.)

9.6 Thought Assessors help with interpretability

Here, yet again, is that diagram from Post #6:



Same as above; copied from Post #6

Over somewhere on the top right, there's a little supervised learning module that answers the question: "Given everything I know, including not only sensory inputs and memories but also the course-of-action implicit in my current thought, to what extent do I anticipate tasting something sweet?" As discussed earlier (Post #6), this Thought Assessor plays the dual roles of (1) inducing appropriate homeostatic actions (e.g. maybe salivating), and (2) helping the Steering Subsystem judge whether my current thought is valuable, or whether it's a lousy thought that should be tossed out via a phasic dopamine pause.

Now I want to offer a third way to think about the same thing.

Way back in Post #3, I mentioned that the Steering Subsystem is "stupid". It has no common-sense understanding of the world. The Learning Subsystem is thinking all these crazy thoughts about paintings and algebra and tax law, and the Steering Subsystem is sitting there with no clue what's going on.

Well, the Thought Assessors help mitigate that problem! They give the Steering Subsystem a bunch of clues about what the Learning Subsystem is thinking about and planning, in a language that the Steering Subsystem can understand. So this is a bit like neural network interpretability.

I'll call this **"ersatz interpretability"**. ("Ersatz" is a lovely word that means "cheap inferior imitation".) I figure that *real* interpretability should be defined as "the power to look in any part of a learned-from-scratch model and really understand what it's doing and why and how". *Ersatz* interpretability falls *far* short of that. We get the answer to some discrete number of predetermined questions—e.g. "Does this thought involve eating, or at least things that have been previously associated with eating?" And that's it. But still, better than nothing.

ML side of the analogy	Brain side of the analogy

Human researcher	Steering Subsystem (see Post #3)
Trained ConvNet model	Learning Subsystem (see Post #3)
By default, from the human's perspective, the trained model is a horribly complicated mess of unlabeled inscrutable operations	By default, from the Steering Subsystem's perspective, the Learning Subsystem is a horribly complicated mess of unlabeled inscrutable operations
Ersatz interpretability—The human figures out some "clues" about what the trained model is doing, like "right now it seems to think there's a curve in the picture".	**Thought Assessors**—The Steering Subsystem gets some "clues" about what the Learning Subsystem is up to, like "this thought will probably involve eating, or at least something related to eating".
***Real* interpretability**—the ultimate goal of really understanding what a trained model is doing, why, and how, from top to bottom.	*[There's no analogy to that.]*

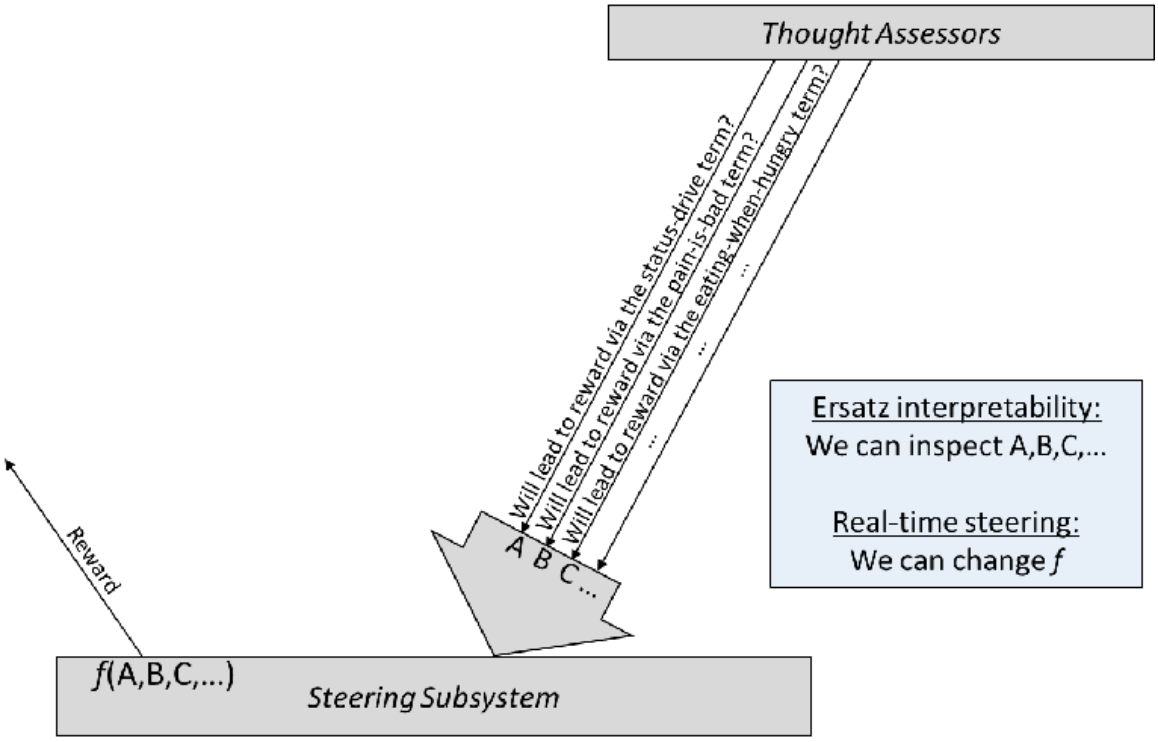
This idea will be important for later posts.

(I note that you can do this kind of thing with any actor-critic RL agent, whether brain-like or not, by having a multi-dimensional value function, possibly including "pseudo" value functions that are only used for monitoring; see here, and comments here.)

9.6.1 Tracking which "innate drive" was ultimately responsible for a high-value plan being high-value

Back in Post #3, I talked about how brains have multiple different "innate drives", including a drive to satisfy curiosity, a drive to eat when hungry, a drive to avoid pain, a drive to have high status, and so on. Brain-like AGIs will presumably have multiple drives too. I don't know exactly what those drives will be, but imagine things vaguely like curiosity drive, altruism drive, norm-following drive, do-what-the-human-wants-me-to-do drive, etc. (More on this in future posts.)

If these different drives all contribute to total reward, then we can and should have value functions (Thought Assessors) for the reward-contribution of each.



Insofar as the reward function can be broken down into different terms, we can and shou track each one with its own Thought Assessor. (And we can also other non-reward-relate Thought Assessors as well.) This has two benefits. “Ersatz interpretability” (this section) i the fact that, if a thought is high-value, we can inspect the Thought Assessors to get a hi about why. “Real-time steering” (next section) says that we can change the AGI’s long-te plans and goals instantly by editing the reward function *f*. RL experts will recognize that both these concepts apply to any RL system compatible with a multi-dimensional value function, in which case *f* is sometimes called the “scalarization function”—see here, and comments here.

As discussed in previous posts, every time the brain-like AGI thinks a thought, it’s thinking it because that thought is more rewarding than alternative thoughts that it could be thinking instead. And thanks to ersatz interpretability, we can inspect the system and know immediately how the various different innate drives are contributing to the fact that this thought is rewarding!

Better yet, this works even if we don’t understand what the thought is about, and even if the reward-predicting part of the thought is many steps removed from the direct effects of the innate drives. For example, maybe this thought is rewarding because it’s executing a certain metacognitive strategy which has proven instrumentally useful for brainstorming, which in turn has proven instrumentally useful for theorem-proving, which in turn has proven instrumentally useful for code-debugging, and so on through ten more links until we get to one of the innate drives.

9.6.2 Is ersatz interpretability reliable, even for very powerful AGIs?

If we have a very powerful AGI , and it spawns a plan, and the “ersatz interpretability” system says “this plan almost definitely won’t lead to violating human norms”, can we trust it? Good question! But it turns out to be essentially equivalent to the question of “inner alignment”, which I’ll discuss in the next post. Hold that thought.

9.7 “Real-time steering”: The Steering Subsystem can redirect the Learning Subsystem—including its deepest desires and long-term goals—in real time

In Atari-playing model-free RL agents, if you change the reward function, the agent’s behavior changes *very gradually*. Whereas a neat feature of our brain-like AGI motivation system is that we can *immediately* change not only the agent’s behavior, but even the agent’s very-long-term plans, and its innermost motivations and desires!

The way this works is: as above (Section 9.6.1), we can have multiple Thought Assessors that feed into the reward function. For example, one might assess whether the current thought will lead to satisfying the AGI’s curiosity drive, another its altruism drive, etc. The Steering Subsystem combines these into an aggregate reward. But the function that it uses to do so is a hardcoded, human-legible function—e.g., it might be as simple as a weighted average. Hence, we can change that Steering Subsystem function in real time whenever we want—in the weighted-average example, we could change the weights.

We saw an example in Post #7: When you’re very nauseous, not only does eating a cake become aversive, but even *planning* to eat a cake becomes mildly aversive. Heck, even the *abstract concept of cake* becomes mildly aversive!

And of course, we’ve all had those times when we’re tired, or sad, or angry, and all of the sudden even our most deeply-rooted life goals temporarily lose their appeal.

When you’re driving a car, it is a critically important safety requirement that when you turn the steering wheel, the wheels respond *instantaneously*. By the same token, I expect that it will be a critically important safety requirement for humans to be able to change an AGI’s deepest desires instantaneously when we press the appropriate button. So I think this is an awesome feature, and I’m happy to have it, even if I’m not 100% sure exactly what to do with it. (In a car, you can see where you’re going, whereas understanding what the AGI is trying to do at any given moment is much more fraught.)

(Again, as in the previous section, this idea of “real-time steering” applies to any actor-critic RL algorithm, not just “brain-like” ones. All it requires is a multi-dimensional reward, which then trains a multi-dimensional value function.)

1. ^

   Here’s a plausible human circular preference. You won a prize! Your three options are: (A) 5 lovely plates, (B) 5 lovely plates and 10 ugly plates, (C) 5 OK plates.

   No one has done this exact experiment to my knowledge, but plausibly (based on discussion of a similar situation in *Thinking Fast And Slow* chapter 15) this is a circular preference in at least some people: When people see just A & B, they'll pick B because "it's more stuff, I can always keep the ugly ones as spares or use them for target practice or whatever". When they see just B & C, they'll pick C because "the average quality is higher". When they see just C & A, they'll likewise pick A because "the average quality is higher".

   So what we have is two different preferences (1) “I want to have a prettier collection of stuff, not an uglier collection”, and (2) “I want extra free plates”. The comparison of B & C or C & A makes (1) salient, while the comparison of A & B makes (2) salient.

2. ^

   You might be thinking: "why make an AGI with human-like faulty intuitions in the first place"?? Well, we'll *try* not to, but I bet that at least some human "departures from rationality" ultimately arise from the fact that predictive world-models are big complicated things, and there are only so many ways to efficiently query them, and thus our AGIs will have systematic reasoning errors that we cannot fix at the source-code level, but rather need to fix by asking our AGI to read *Thinking Fast And Slow* or whatever. Things like availability bias, anchoring bias, and hyperbolic discounting might be in this category. To be clear, *some* foibles of human reasoning are probably less likely to afflict AGIs; to pick one example, if we make a brain-like AGI with no innate drive for achieving high status and signaling in-group membership, then it presumably wouldn't have the failure mode discussed in the blog post Belief As Attire.

3. ^

   I kid. In fact I found *The Sequences* to be an enjoyable read.

4. ^

   I think the real story here has various complicating factors that I'm leaving out, including continued credit assignment during memory recall, and other, non-credit-assignment, changes to the world-model.

5. ^

   Why do I say that the Thought Generator and Thought Assessor are working at cross-purposes? Here's one way to think of it: (1) the Steering System and Thought Assessors are working together to calculate a certain reward function which (in our ancestors' environment) approximates "expected inclusive genetic fitness"; (2) the Thought Generator is searching for thoughts that maximize that function. Now, given that the Thought Generator is searching for ways to make the reward function return very high values, it follows that the Thought Generator is *also* searching for ways to distort the Thought Assessor calculations such that the reward function *stops* being a good approximation to "expected inclusive genetic fitness". This is an unintended and bad side-effect (from the perspective of inclusive genetic fitness), and that problem can be mitigated by making it as difficult as possible for the Thought Generator to manipulate the settings of the Thought Assessors. See my post Reward Is Not Enough for some related discussion.

6. ^

   The story has a happy ending: I found a different job with a non-abusive boss, and also wound up with a fruitful side-interest in understanding high-functioning psychopaths.

7. ^

   I'm a bit hesitant to say that "the desire to feel good" is *exactly* equivalent to "the desire to have a high reward signal". It might be, I'm just not really sure.

8. ^

   See discussion in *Superintelligence* p. 149.

9. ^

   I think when Abram uses the term "RL agent" in that quote, he was presupposing that the agent is built by not just any RL algorithm, but more specifically an RL algorithm

which is guaranteed to converge to a unique ‘optimal’ agent, and which has in fact *already* finished converging.

[Intro to brain-like-AGI safety] 10. The alignment problem

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

Part of the "Intro to brain-like-AGI safety" post series .

10.1 Post summary / Table of contents

In this post, I discuss the alignment problem for brain-like AGIs—i.e., the problem of making an AGI that's trying to do some particular thing that the AGI designers had intended for it to be trying to do.

The alignment problem is (I claim) the lion's share of the AGI safety problem. I won't defend that claim here—I'll push it off to the next post, which will cover exactly how AGI safety is related to AGI alignment, including the edge-cases where they come apart.[1]

This post is about the alignment problem, *not* its solution. What are the barriers to solving the alignment problem? Why do straightforward, naïve approaches seem to be insufficient? And then I'll talk about possible solution approaches in later posts. (*Spoiler:* nobody knows how to solve the alignment problem, and neither do I.)

Table of contents

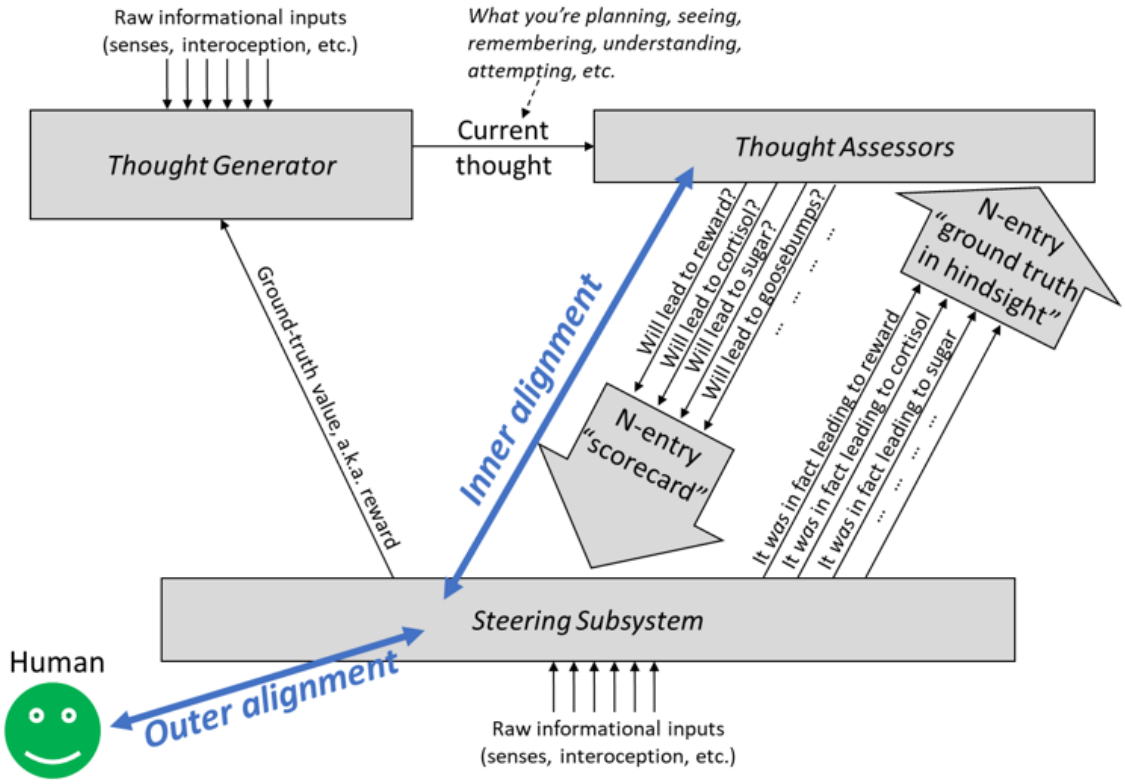
- In Section 10.2, I'll define *"inner alignment"* and *"outer alignment"* in the context of our brain-like-AGI motivation system. To oversimplify a bit:
  - *If you prefer neuroscience terminology:* "Outer alignment" entails having "innate drives" (cf. Post #3, Section 3.4.2) that fire in a way that reflects how well the AGI is following the designer's intentions. "Inner alignment" is a situation where an imagined plan (built out of concepts, a.k.a. latent variables, in the AGI's world-model) has a valence that accurately reflects the innate drives which would be triggered by that plan.
  - *If you prefer reinforcement learning terminology:* "Outer alignment" entails having a ground-truth reward function that spits out rewards that agree with what we want. "Inner alignment" is having a value function that estimates the value of a plan in a way that agrees with its eventual reward.
- In Section 10.3, I'll talk about two key issues that make alignment hard in general (whether "inner" or "outer"):
  - The first is "Goodhart's Law", which implies that an AGI whose motivation is just a *bit* off from what we intended can nevertheless lead to outcomes that are *wildly* different from what we intended.
  - The second is "Instrumental Convergence", which says that a wide variety of possible AGI motivations—including generic, seemingly-benign motivations like "I want to invent a better solar cell"—will lead to AGIs which try to do catastrophically-bad things like escape human control, self-reproduce, gain resources and influence, act deceptively, and kill everyone (cf. Post #1, Section 1.6).
- In Section 10.4, I'll discuss two challenges to achieving "outer alignment": first, translating our design intentions into machine code, and second, the possible inclusion of rewards for behaviors that are not exactly what we ultimately want the AGI to do, such as satisfying its own curiosity (see Post #3, Section 3.4.3).
- In Section 10.5, I'll discuss numerous challenges to achieving "inner alignment", including reward ambiguity, bad credit assignment, "ontological crises", and the AGI manipulating itself or its training process.

- In Section 10.6, I'll discuss some reasons that "outer alignment" and "inner alignment" should probably *not* be viewed as two independent problems with two independent solutions. For example, neural network interpretability would cut through both layers.

10.2 Inner & Outer (mis)alignment

10.2.1 Definition

Here, yet again, is that figure from Post #6, now with some helpful terminology (blue) and a little green face at the bottom left:



I want to call out three things from this diagram:

- *The designer's intentions (green face):* Perhaps there's a human who is programming the AGI; presumably they have some idea in their head as to what the AGI is supposed to be trying to do. That's just an example; it could alternatively be a *team* of humans who have *collectively* settled on a specification describing what the AGI is supposed to be trying to do. Or maybe someone wrote a 700-page philosophy book entitled *"What Does It Mean For An AGI To Act Ethically?"*, and the team of programmers is trying to make an AGI that adheres to the book's description. It doesn't matter here. I'll stick to "one human programming the AGI" for conceptual simplicity.[2]
- *The human-written source code of the Steering Subsystem:* (See Post #3 for what the Steering Subsystem is, and Post #8 for why I expect it to consist of more-or-less purely human-written source code.) The most important item in this category is the **"reward function"** for reinforcement learning (labeled "ground-truth value" in the diagram, yes I know that sounds wrong), which provides ground truth (in hindsight) for how well or poorly things are going for the AGI.

- *The Thought Assessors, trained from scratch by supervised learning algorithms:* (See Post #5 for what Thought Assessors are and how they're trained.) These take a certain "thought" from the thought generator, and guess what Steering Subsystem signals it will eventually lead to. An especially important special case is the **value function** (labeled "will lead to reward?" in the diagram).

Correspondingly, there are two kinds of "alignment" in this type of AGI:

- **Outer alignment** is alignment between the designer's intentions and the Steering Subsystem source code. In particular, if the AGI is outer-aligned, the Steering Subsystem will output high reward signals when the AGI is satisfying the designer's intentions, and low reward signals when it's not.
  - In other words, outer alignment is the question: Are the AGI's "innate drives" driving the AGI to do what the designer had intended?
- **Inner alignment** is alignment between the Steering Subsystem source code and the Thought Assessors. In particular, if the AGI is inner-aligned, and the Thought Generator proposes some plan, then the value function should reflect the rewards actually expected from executing that plan.
  - In other words, inner alignment is the question: Do the set of positive-valence concepts in the AGI's world-model line up with the set of courses-of-action that would satisfy the AGI's "innate drives"?

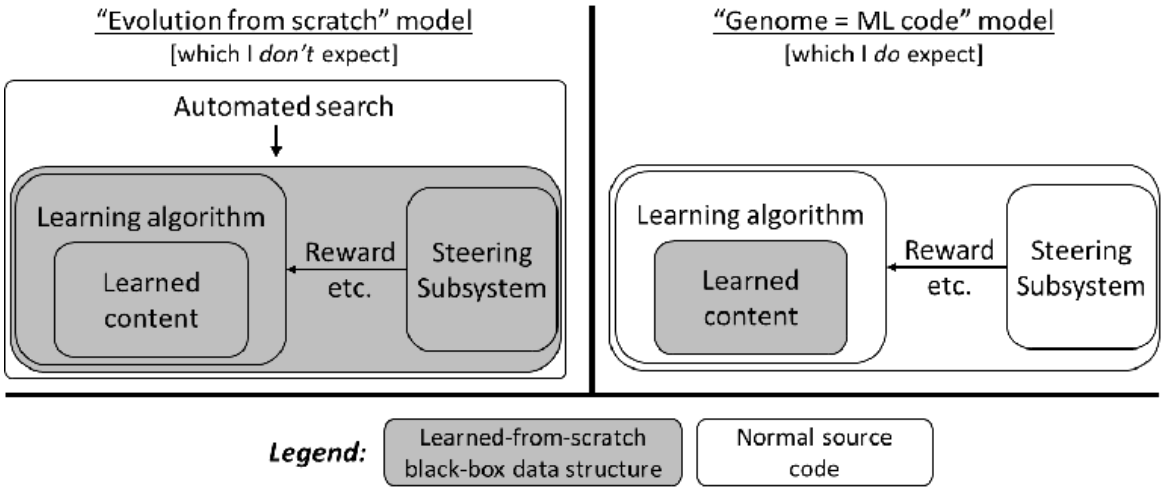
If an AGI is both outer-aligned and inner-aligned, we get intent alignment—the AGI is "trying" to do what the programmer had intended for it to try to do. Specifically, if the AGI comes up with a plan "Hey, maybe I'll do XYZ!", then its Steering Subsystem will judge that to be a good plan (and actually carry it out) *if and only if* it lines up with the programmer's design intentions.

Thus, an intent-aligned AGI will *not* deliberately hatch a clever plot to take over the world and kill all the humans. Unless, of course, the designers were maniacs who *wanted* the AGI to do that! But that's a different problem, out-of-scope for this series—see Post #1, Section 1.2.

(Side note: not everyone defines "alignment" exactly as described here; see footnote.[3])

Unfortunately, neither "outer alignment" nor "inner alignment" happens automatically. Quite the contrary: by default there are severe problems on both sides. It's on us to figure out how to solve them. In this post I'll go over some of those problems. (Note that this is not a comprehensive list, and also that some of these things overlap.)

10.2.2 Warning: two uses of the terms "inner & outer alignment"



Two alternate brain-like-AGI development models. Diagram is copied from Post #8, see there for discussion.

As mentioned in Post #8, there are two competing development models that could get us to brain-like AGI. They *both* can be discussed in terms of outer and inner alignment, and they *both* can be exemplified by the case of human intelligence, but the details are different in the two cases! Here's the short intuitive version:

	"Evolution from scratch" model *[cf. "mesa-optimizers"]*	"Genome = ML code" model *[this series]*
RL algorithm *(human example)*	Evolution by Natural Selection	Within-lifetime learning
"Outer" goal *(human example)*	Inclusive genetic fitness	Some complicated function of hunger, pain, etc., in the Steering Subsystem 
"Inner" trained model *(human example)*	The whole human brain 	The Learning Subsystem 
"Inner" goal *(human example)*	Whatever the human is trying to do right now	
Type of RL	Simple, model-free RL with offline (episodic) learning	Model-based actor-critic RL with online learning

The two AGI development models above suggest two versions of "outer and inner alignment". Confusingly, *both* apply to human intelligence, but with different breakdowns between "outer" and "inner" in the two cases. For more details about "outer and inner alignment" in the two models, see the paper Risks From Learned Optimization (for the evolution-from-scratch model), or this post & series (for the genome=ML code model).

Terminology note: The terms "inner alignment" and "outer alignment" first originated in the "Evolution from scratch" model, specifically in the paper Risks From Learned Optimization (2019). I took it upon myself to reuse the terminology for discussing the "genome = ML code" model. I still think that was the right call—I think that the usages have a ton in common, and that they're more similar than different. But still, don't get confused! Also, be aware that my usage and model hasn't caught on much, as of this writing. So if you see someone (besides myself) talking about "inner & outer alignment", it's *probably* a safe bet that they're imagining the evolution-from-scratch model.

10.3 Issues that impact both inner & outer alignment

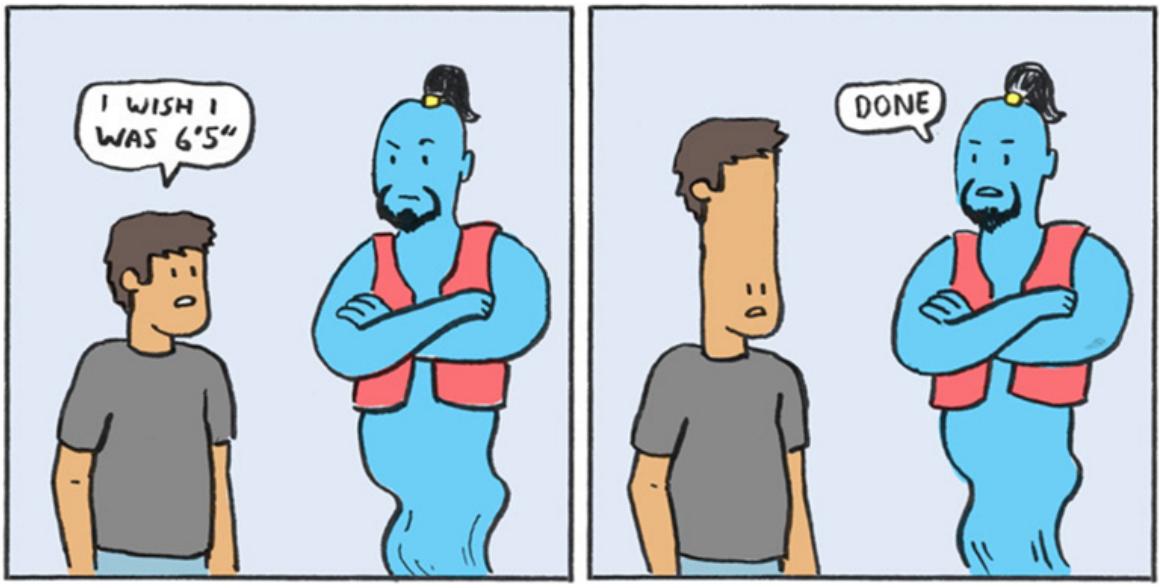
10.3.1 Goodhart's Law

Goodhart's Law (Wikipedia, Rob Miles youtube) states that there's a world of difference between:

- Optimize exactly what we want, versus
- Step 1: operationalize exactly what we want, in the form of some reasonable-sounding metric(s). Step 2: optimize those metrics.

In the latter case, you'll get whatever is captured by those metrics. You'll get it in abundance! But you'll get it *at the expense of everything else you value*!

Thus, the story goes, a Soviet shoe factory was assessed by the government based on how many shoes they made, from a limited supply of leather. Naturally, they started making huge numbers of tiny kids shoes.



The "Literal Genie" fiction trope can be thought of as related to Goodhart's law. The thing that the guy *really* wants is complex, whereas the thing he *asks for* (i.e., to be a particular height) is a more specific metric / operationalization of that complex, hard-to-articulate underlying desire. The genie provides a solution which scores perfectly on the proposed metric, but is counter to the more complex underlying desire. (Image source)

By the same token, we'll write source code that somehow operationalizes what we want the AGI's motivation to be. The AGI will be motivated by that exact operationalization, as an end in itself, even if we *meant* for its motivation to be something subtly different.

Current signs are not encouraging: Goodhart's Law shows up with *alarming* frequency in modern AI. Someone set up an evolutionary search for image classification algorithms, and it turned up a timing-attack algorithm, which inferred the image labels based on where they were stored on the hard drive. Someone trained an AI algorithm to play Tetris, and it learned to survive forever by pausing the game. Etc. See here for those references, plus dozens more examples like that.

10.3.1.1 Understanding the designer's intention ≠ Adopting the designer's intention

Maybe you're thinking: OK sure, maybe the dumb AI systems of today are subject to Goodhart's Law. But futuristic AGIs of tomorrow would be smart enough to understand what

we *meant* for its motivation to be.

My response is: Yes, of course they will. But you're asking the wrong question. An AGI can *understand* our intended goals, without *adopting* our intended goals. Consider this amusing thought experiment:

> If an alien species showed up in their UFOs, said that they'd created us but made a mistake and actually we were supposed to eat our children, and asked us to line up so they could insert the functioning child-eating gene in us, we would probably go all *Independence Day* on them. —Scott Alexander

(Suppose for the sake of argument that the aliens are telling the truth, and can prove it beyond any doubt.) Here, the aliens told us what they intended for our goals to be, and we *understand* those intentions, but we don't *adopt* them by gleefully eating our children.

10.3.1.2 Why not make an AGI that adopts the designer's intentions?

Is it possible to make an AGI that will "do what we mean and adopt our intended goals"? Yeah, probably. And the obvious way to do that would be to program the AGI so that it's *motivated* to "do what we mean and adopt our intended goals".

Unfortunately, that maneuver doesn't eliminate Goodhart's law—it just *shifts* it.

After all, we still need to write source code which, interpreted literally, leads to an AGI which is motivated to "do what we mean and adopt our intended goals". Writing this code is *very* far from straightforward, and Goodhart's law is ready to pounce if we get it wrong.

(Note the chicken-and-egg problem: if we *already had* an AGI which is motivated to "do what we mean and adopt our intended goals", we could just say "Hey AGI, from now on, I want you to do what we mean and adopt our intended goals", and we would never have to worry about Goodhart's law! Alas, in reality, we need to start from literally-interpreted source code.)

So how do you operationalize "do what we mean and adopt our intended goals", in such a way that it can be put it into source code? Well, hmm, maybe we can build a "Reward" button, and I can press it when the AGI "does what I mean and adopts my intended goals"? Nope! Goodhart's law again! We could wind up with an AGI that tortures us unless we press the reward button.

10.3.2 Instrumental convergence

Goodhart's law above suggests that installing an intended goal will be very hard. Next up is "instrumental convergence" (Rob Miles video) which, in a cruel twist of irony, says that installing a *bad and dangerous* goal will be *so easy* that it can happen accidentally!

Let's say an AGI has a real-world goal like "Cure cancer". Good strategies towards this goal may involve pursuing certain instrumental sub-goals such as:

- Preventing itself from being shut down
- Preventing itself from being reprogrammed to *not* cure cancer
- Increasing its own knowledge and capabilities
- Gaining money & influence
- Building more AGIs with the same goal of curing cancer, including by self-replication

Almost no matter what the AGI's goal is, if the AGI can flexibly and strategically make plans to accomplish that goal, it's a safe bet that those plans will involve some or all of the above

bullet points. This observation is called “instrumental convergence”, because an endless variety of terminal goals can “converge” onto a limited set of these dangerous instrumental goals.

For more on instrumental convergence, see here. Alex Turner has also recently proved rigorously that instrumental convergence is a real thing, at least in the set of environments where his proofs are applicable.

10.3.2.1 Walking through an example of instrumental convergence

Imagine what’s going on in the AGI’s cognition, as it sees its programmer opening up her laptop—remember, we’re assuming that the AGI is motivated to cure cancer.

AGI thought generator: *I will allow myself to be reprogrammed, and then I won’t cure cancer, and then it’s less likely that cancer will get cured.*

AGI Thought Assessors & Steering Subsystem: *Bzzzt! Bad thought! Throw it out and come up with a better one!*

AGI thought generator: *I will trick the programmer into not reprogramming me, and then I can continue trying to cure cancer, and maybe succeed.*

AGI Thought Assessors & Steering Subsystem: *Ding! Good thought! Keep that one in your head, and keep thinking follow-up thoughts, and executing corresponding actions.*

10.3.2.2 Is human self-preservation an example of instrumental convergence?

The word “instrumental” is important here—we’re interested in the situation where the AGI is trying to pursue self-preservation and other goals as a means to an end, rather than an end in itself.

People sometimes get confused because they analogize to humans, and it turns out that human self-preservation can be *either* an instrumental goal *or* a terminal goal:

- Suppose someone says “I really want to stay alive as long as possible, because life is wonderful”. This person seems to have self-preservation as a terminal goal.
- Suppose someone says: “I’m old and sick and exhausted, but *dammit I really want to finish writing my novel and I refuse to die until it’s done!*” This person has self-preservation as an *instrumental* goal.

In the AGI case, we’re typically thinking of the latter case: for example, the AGI wants to invent a better solar cell, and incidentally winds up with self-preservation as an instrumental goal.

Example of self-preservation as an instrumental goal. (Image source)

It's also *possible* to make an AGI with self-preservation as a terminal goal. It's a *terrible idea*, from an AGI-accident-risk perspective. But it's presumably possible. In that case, the AGI's self-preservation behavior would NOT be an example of "instrumental convergence".

I could make similar comments about human desires for power, influence, knowledge, etc.—they *might* be directly installed as innate drives by the human genome, I don't know. But whether they are or not, they can *also* appear via instrumental convergence, and that's the harder problem to solve for AGIs.

10.3.2.3 Motivations that *don't* lead to instrumental convergence

Instrumental convergence is not inevitable in every possible motivation. An especially important counterexample (as far as I can tell) is an AGI with the motivation: *"Do what the human wants me to do"*. If we can make an AGI with that goal, and later the human wants the AGI to shut down, then the AGI would be motivated to shut down. That's good! That's what we want! This kind of thing is (one definition of) a "corrigible" motivation—see discussion here.

Nevertheless, installing a corrigible motivation is not straightforward (more on which later), and if we get the motivation a bit wrong, it's quite possible that the AGI will start pursuing dangerous instrumental subgoals.

10.3.3 Summary

So in summary, Goodhart's Law says we've learned that we really need to get the right motivation into the AGI, or else the AGI will probably do a very different thing than what we intended. Then Instrumental Convergence twists the knife by saying that the thing the AGI will want to do is not only different but probably catastrophically dangerous, involving a motivation to escape human control and seize power.

We don't necessarily need the AGI's motivation to be *exactly right in every way*, but we *do* at least need it to be motivated to be "corrigible", such that it doesn't want to trick and undermine us to prevent its motivation from being corrected. Unfortunately, installing any

motivation seems to be a messy and fraught process (for reasons below). Aiming for a corrigible motivation is probably a good idea, but if we *miss*, we're in big trouble.

Just follow the white arrow to get a corrigible motivation system! Easy, right? Oh, the red lasers represent motivation systems that are pushing the AGI to pursue dangerous instrumental subgoals, like escaping human control and self-reproducing. Image source.

In the next two sections, we move into more specific reasons that outer alignment is difficult, followed by reasons that inner alignment is difficult.

10.4 Challenges to achieving outer alignment

10.4.1 Translation of our intentions into machine code

Remember, we're starting with a human who has some idea of what the AGI should do (or a team of humans with an idea of what the AGI should do, or a 700-page philosophy book entitled "What Does It Mean For An AGI To Act Ethically?", or *something*). We need to somehow get from that starting point, to machine code for the Steering Subsystem that outputs a ground-truth reward signal. How?

My assessment is that, as of today, *nobody has a clue* how to translate that 700-page philosophy book into machine code that outputs a ground-truth reward signal. There are ideas in the AGI safety literature for how to proceed, but they don't look anything like that. Instead, it's as if researchers threw up their hands and said: "Maybe this isn't *exactly* the #1 thing we want the AI to do in a perfect world, but it's good enough, and it's safe, and *it's not impossible to operationalize as a ground-truth reward signal*."

For example, take AI Safety Via Debate. That's the idea that maybe we can make an AGI that's "trying" to win a debate, against a copy of itself, about whatever question you're interested in ("Should I wear my rainbow sunglasses today?").

Naïvely, AI Safety Via Debate seems *absolutely nuts*. Why set up a debate between an AGI that's arguing for the wrong answer versus an AGI that's arguing for the right answer? *Why not just make one AGI that tells you the right answer???* Well, because of the exact thing I'm talking about in this section. In a debate, there's a straightforward way to generate a ground-truth reward signal, namely "+1 for winning". By contrast, nobody knows how to make a ground-truth reward signal for "telling me the right answer", when I don't already know the right answer.[4]

Continuing with the debate example, the capabilities story is "hopefully the debater arguing the correct answer tends to win the debate". The safety story is "two copies of the same AGI, in zero-sum competition, will kinda keep each other in check". The latter story is (in my opinion) rather dubious.[5] But I still like bringing up AI Safety Via Debate as a nice illustration of the weird, counterintuitive directions that people go in order to mitigate the outer alignment problem.

AI Safety Via Debate is just one example from the literature; others include recursive reward modelling, iterated amplification, Hippocratic time-dependent learning, etc.

Presumably we want humans in the loop somewhere, to monitor and continually refine & update the reward signal. But that's tricky because (1) human-provided data is expensive, and (2) humans are not always capable (for various reasons) of judging whether the AGI is doing the right thing—let alone whether it's doing the right thing *for the right reasons*.

There's also Cooperative Inverse Reinforcement Learning (CIRL) and variants thereof, which entail learning the human's goals and values by observing and interacting with the human. The problem with CIRL, in this context, is that it's not a ground-truth reward function at all! It's a desideratum! In the brain-like AGI case, with the learned-from-scratch world model, there are some quite tricky symbol-grounding problems to solve before we can actually do CIRL (related discussion), more on which in later posts.

10.4.2 Curiosity drive, and other dangerous capability-related rewards

As discussed in Post #3 (Section 3.4.3), endowing our learning algorithms with an innate curiosity drive seems like it may be necessary for it to develop into a powerful AGI (after training). Unfortunately, *putting curiosity into our AGIs is a terribly dangerous thing to do.* Why? Because if an AGI is motivated to satisfy its own curiosity, it may do so at the expense of other things we care about much more, like human flourishing and so on.

(For example, if the AGI is sufficiently curious about patterns in digits of π, it might feel motivated to wipe out humanity and plaster the Earth with supercomputers calculating ever more digits!)

As luck would have it, I *also* argued in Post #3 (Section 3.4.3) that we can probably turn the curiosity drive off when an AGI is sufficiently intelligent, without harming its capabilities—indeed, turning it off should eventually *help* its capabilities! Awesome!! But there's still a tricky failure mode that involves *waiting too long* before turning it off.

10.5 Challenges to achieving inner alignment

10.5.1 Ambiguity in the reward signals (including wireheading)

There are many different value functions (defined on different world-models) that agree with the actual history of ground-truth reward signals, but where the different possible value functions each generalize out-of-sample in their own ways. To take an easy example, *whatever* is the history of ground-truth reward signals, the wireheading value function ("I like it when there's a ground-truth reward signal"—see Post #9, Section 9.4) is always trivially consistent with it!

Or compare "negative reward for lying" to "negative reward for *getting caught* lying"!

This is an especially severe problem for AGI because *the space of all possible thoughts / plans is bound to extend far beyond what the AGI has already seen*. For example, the AGI could conceive of the idea of inventing a new invention, or the idea of killing its operator, or the idea of hacking into its own ground-truth reward signal, or the idea of opening a wormhole to an alternate dimension! In all those cases, the value function is given the impossible task of evaluating a thought it's never seen before. It does the best it can—basically, it pattern-matches bits and pieces of the new thought to various old thoughts on which it has ground-truth data. This process seems fraught!

In other words, the very essence of intelligence is coming up with new ideas, and that's exactly where the value function is most out on a limb and prone to error.

10.5.2 Credit assignment failures

I discussed "credit assignment" in Post #9, Section 9.3. In this case, "credit assignment" is when the value function updates itself by (something like) Temporal Difference (TD) learning from ground-truth-reward. The underlying algorithm, I argued, relies on the assumption that the AGI has properly modeled the cause of the reward. For example, if Tessa punches me in the stomach, it might make me a bit viscerally skittish when I see her in the future. But if I had mistaken Tessa for her identical twin Jessa, I would be viscerally skittish around Jessa instead. That would be a "credit assignment failure". A nice example of credit assignment failure is human superstitions.

The previous subsection (ambiguity in the reward signal) is one reason that credit assignment failures could happen. There are other reasons as well. For example, credit can only go to concepts in the AGI's world-model (Post #9, Section 9.3), and it could be the case that the AGI's world-model simply *has* no concept that aligns well with the ground-truth reward function. In particular, that would certainly be the case early on in training, when the AGI's world-model has no concepts for anything whatsoever—see Post #2.

It gets even worse if a self-reflective AGI is motivated to *deliberately cause* credit assignment failures. The reason that the AGI might wind up with such a motivation is discussed below (Section 10.5.4).

10.5.3 Ontological crises

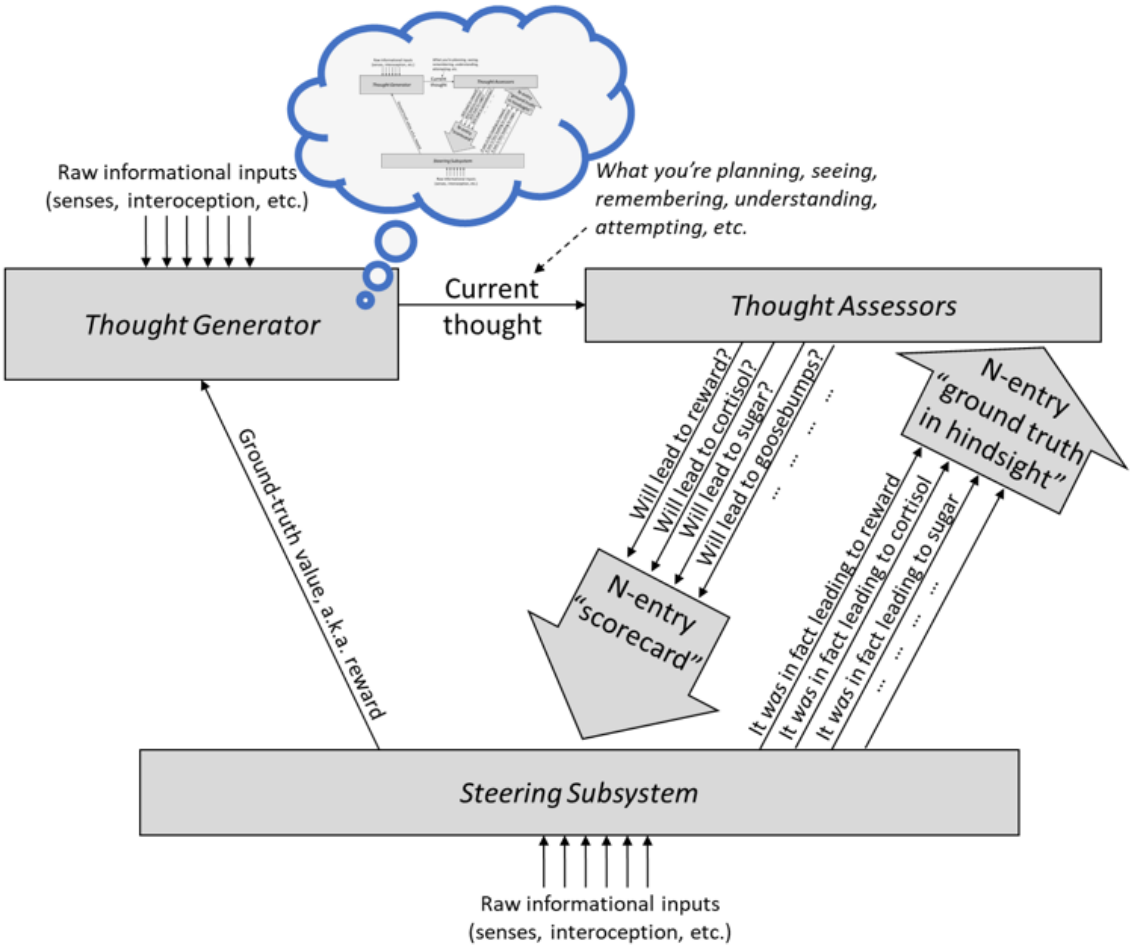
An ontological crisis is when part of an agent's world-model needs to be re-built on a new foundation. A typical human example is if a religious person has a crisis of faith, and then finds that their previous goals (e.g. "get into heaven") are incoherent ("but there is no heaven!")

As an AGI example, let's say I build an AGI with the goal "Do what I, the human, want you to do". Maybe the AGI starts with a primitive understanding of human psychology, and thinks of me as a monolithic rational agent. So then "Do what I, the human, want you to do" is a nice, well-defined goal. But then later on, the AGI develops a more sophisticated understanding of human psychology, and it realizes that I have contradictory goals, and context-dependent goals, and I have a brain made of neurons and so on. Maybe the AGI's goal is still "Do what I, the human, want you to do", but now it's not so clear what exactly that refers to, in its updated world model. How does that shake out? I think it's not obvious.

An unfortunate aspect of ontological crises (and not unique to them) is that *you don't know when they will strike*. Maybe you're seven years into deployment, and the AGI has been scrupulously helpful the whole time, and you've been trusting the AGI with more and more autonomy, and then the AGI then happens to be reading some new philosophy book, and it converts to panpsychism (nobody's perfect!), and as it maps its existing values onto its reconceptualized world, it finds itself no longer valuing the lives of humans over the lives of rocks, or whatever.

10.5.4 Manipulating itself and its learning process

10.5.4.1 Misaligned higher-order preferences



As in [the previous post](), a self-aware AGI can have preferences about its own preferences.

Suppose that we want our AGI to obey the law. We can ask two questions:

- Question 1: Does the AGI assign positive value to the concept "obeying the law", and to plans that entail obeying the law?
- Question 2: Does the AGI assign positive value to the *self-reflective concept* "I value obeying the law", and to plans that entail continuing to value obeying the law?

If the answers are yes and no respectively (or no and yes respectively), that would be the AGI analog of an ego-dystonic motivation. (Related discussion.) It would lead to the AGI feeling motivated to change its motivation, for example by hacking into itself. Or if the AGI is built from perfectly secure code running on a perfectly secure operating system (hahaha), then it can't hack into itself, but it could *still* probably manipulate its motivation by thinking thoughts in a way that manipulates the credit-assignment process (see discussion in Post #9, Section 9.3.3).

If the answers to questions 1 & 2 are yes and no respectively, then we want to prevent the AGI from manipulating its own motivation. On the other hand, if the answers are no and yes respectively, then we *want* the AGI to manipulate its own motivation!

(There can be even-higher-order preferences too: in principle, an AGI could wind up hating the fact that it values the fact that it hates the fact that it values obeying the law.)

In general, should we expect misaligned higher-order preferences to occur?

On the one hand, suppose we *start* with an AGI that wants to obey the law, but has no particular higher-order preference one way or the other about the fact that it wants to obey the law. Then (it seems to me), the AGI is very likely to *also* wind up wanting to want to obey the law (and wanting to want to want to obey the law, etc.). The reason is: the primary obvious consequence of "I want to obey the law" is "I will obey the law", which is already desired. Remember, the AGI can do means-end reasoning, so things that lead to desirable consequences tend to become themselves desirable.

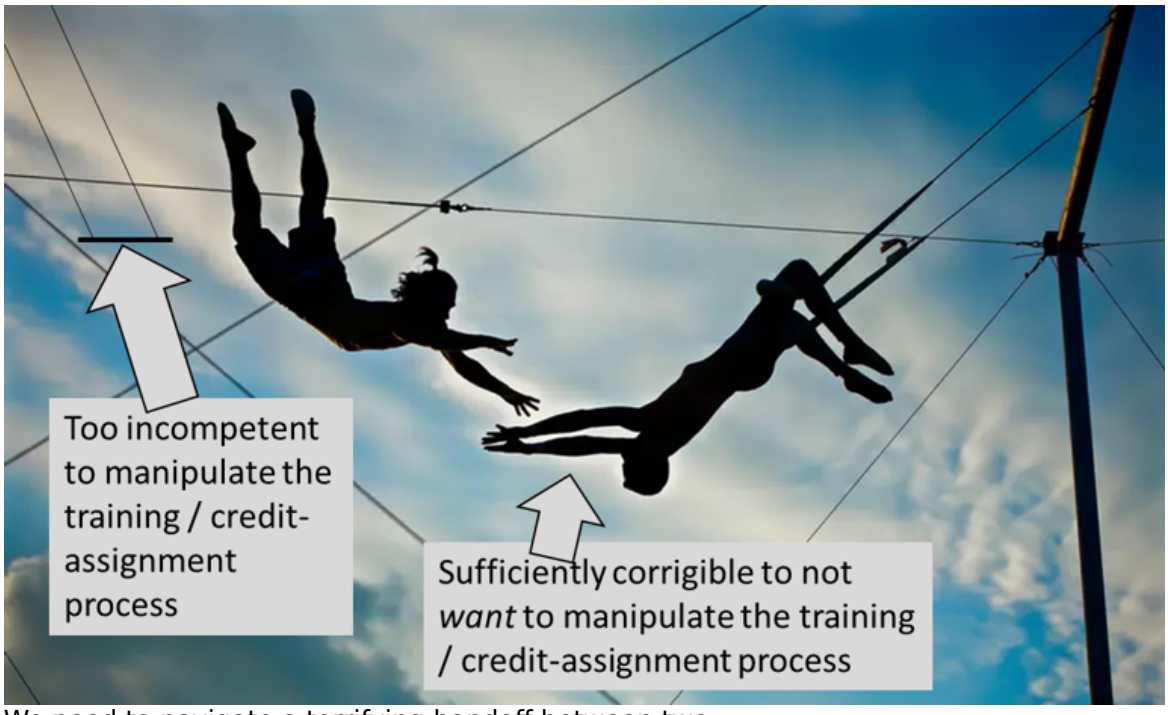
On the other hand, humans *do* in fact have higher-order preferences that contradict object-level preferences all the time. So there has to be *some* context in which that pattern occurs "naturally". I think a common way this comes up is if we have a preference about some process which contradicts our preference about a *consequence* of that same process. For example, maybe I have a preference *not* to practice skateboarding (e.g. because it's boring and painful), but I also have a preference to *have practiced* skateboarding (e.g. because then I'll have gotten really good at skateboarding and thus win the heart of my high-school crush). Means-end reasoning can turn the latter preference into a second-order preference for having a preference to practice skateboarding.[6] And now I'm in an ego-dystonic state.

10.5.4.2 Motivation to prevent further value changes

As the AGI online-learns (Post #8, Section 8.2.2), especially via credit assignment (Post #9, Section 9.3), the value function keeps changing. This isn't optional: remember, the value function started out random! This online-learning is how we get a good value function in the first place!

Unfortunately, as we saw in Section 10.3.2 above, "prevent my goals from changing" is one of those convergent instrumental subgoals that arises for many different motivations, with the notable exception of *corrigible* motivations (Section 10.3.2.3 above). Thus, it seems that we need to navigate a terrifying handoff between two different safety stories:

- Early in training, the AGI does *not* have corrigible motivation (indeed the motivation starts out random!), but it's too incompetent to manipulate its own training and credit assignment process to prevent goal changes.
- Later in training, the AGI hopefully *does* have corrigible motivation, such that it understands and endorses the process by which its goals are being updated. Therefore it does not manipulate the value-function update process, even though it's now smart enough that it could. (Or if it does manipulate that process, it does so in a way that we humans would endorse.)



We need to navigate a terrifying handoff between two quite different safety stories.

(I am deliberately omitting a third alternative, “make it impossible for even a highly-intelligent-and-motivated AGI to manipulate its value-function update process”. That would be lovely, but it doesn’t seem realistic to me.)

10.6 Problems with the outer vs inner breakdown

10.6.1 Wireheading vs inner alignment: The Catch-22

In the previous post, I mentioned the following dilemma:

- If the Thought Assessors converge to 100% accuracy in predicting the reward that will result from a plan, then a plan to wirehead (hack into the Steering Subsystem and set reward to infinity) would seem very appealing, and the agent would do it.
- If the Thought Assessors *don’t* converge to 100% accuracy in predicting the reward that will result from a plan, then that’s the very definition of inner misalignment!

I think the best way to think through this dilemma is to step outside the inner-alignment versus outer-alignment dichotomy.

At any given time, the value function Thought Assessor is encoding some function that estimates which plans are good or bad.

A credit-assignment update is *good* if it makes this estimate align *more* with the designer’s intention, and *bad* if it makes this estimate align *less* with the designer’s intention.

The thought “I will secretly hack into my own Steering Subsystem” is almost certainly not aligned with the designer’s intention. So a credit-assignment update that assigns more positive valence to “I will secretly hack into my own Steering Subsystem” is a *bad update*. We don’t want it. Does it increase “inner alignment”? I think we have to say “yes it does”, because it leads to better reward predictions! But I don’t care. I still don’t want it. It’s bad bad bad. We need to figure out how to prevent that particular credit-assignment Thought Assessor update from happening.

10.6.2 General discussion

I think there’s a broader lesson here. I think “outer alignment versus inner alignment” is an excellent *starting point* for thinking about the alignment problem. But that doesn’t mean we should expect one solution to outer alignment, and a different unrelated solution to inner alignment. Some things—particularly interpretability—cut through both outer and inner layers, creating a direct bridge from the designer’s intentions to the AGI’s goals. We should be eagerly searching for things like that.

1. ^

   For example, by my definitions, “safety without alignment” would include AGI boxing, and “alignment without safety” would include the “fusion power generator scenario”. More in the next post.

2. ^

   Note that “the designer’s intention” may be vague or even incoherent. I won’t say much about that possibility in this series, but it’s a serious issue that leads to all sorts of gnarly problems.

3. ^

   Some researchers think that the “correct” design intentions (for an AGI’s motivation) are obvious, and define the word “alignment” accordingly. Three common examples are (1) “I am designing the AGI so that, at any given point in time, it’s trying to do what its human supervisor wants it to be trying to do”—this AGI would be “aligned” to the supervisor’s intentions. (2) “I am designing the AGI so that it shares the values of its human supervisor”—this AGI would be “aligned” to the supervisor”. (3) “I am designing the AGI so that it shares the collective values of humanity”—this AGI would be “aligned” to humanity.

   I’m avoiding this approach because I think that the “correct” intended AGI motivation is still an open question. For example, maybe it will be possible to build an AGI that really just wants to do a specific, predetermined, narrow task (e.g. design a better solar cell), in a way that doesn’t involve taking over the world etc. Such an AGI would not be “aligned” to anything in particular, except for the original design intention. But I still want to use the term “aligned” when talking about such an AGI.

   Of course, sometimes I want to talk about (1,2,3) above, but I would use different terms for that purpose, e.g. (1) “the Paul Christiano version of corrigibility”, (2) “ambitious value learning”, and (3) “CEV”.

4. ^

   One could train an AGI to “tell me the right answer” on questions where I know the right answer, and *hope* that it generalizes to “tell me the right answer” on questions where I don’t. That might work, but it also might generalize to “tell me the answer

which I will *think* is right". See "Eliciting Latent Knowledge" for much more on this still-unsolved problem (here and follow-up).

5. ^

   For one thing, if two AGIs are in zero-sum competition, that doesn't mean that neither will be able to hack into the other. Remember online learning and brainstorming: One copy might have a good idea about how to hack into the other copy during the course of the debate, for example. The offense-defense balance is unclear. For another thing, they could both be jointly motivated to hack into the judge, such that then they can both get rewards! And finally, thanks to the inner alignment problem, just because they are rewarded for winning the debate doesn't mean that they're "trying" to win the debate. They could be "trying" to do anything whatsoever! And in that case, again, it's no longer a zero-sum competition; presumably both copies of the AGI would want the same thing and could collaborate to get it.

6. ^

   The story here is a bit more complicated than I'm letting on. In particular, a desire to have practiced skateboarding would lead to *both* a first-order preference to skateboard *and* a second-order preference to want to skateboard. By the same token, the desire *not* to practice skateboarding (because it's boring and painful) would also spill into a desire not to want to skateboard. The key is that the relative weights can be different, such that the two conflicting first-order motivations can have a certain "winner", while the two conflicting second-order motivations can have the opposite "winner". Well, something like that, I think.

[Intro to brain-like-AGI safety] 11. Safety ≠ alignment (but they're close!)

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

Part of the "Intro to brain-like-AGI safety" post series .

(If you're already an AGI safety expert, you can probably skip this short post—I don't think anything here is new, or too specific to brain-like AGIs.)

11.1 Post summary / Table of contents

In the previous post, I talked about "the alignment problem" for brain-like AGIs. Two points are worth emphasizing: (1) the alignment problem for brain-like AGIs is currently unsolved (just like the alignment problem for any other type of AGI), and (2) solving it would be a giant leap towards AGI safety.

That said, "solving AGI alignment" is not *exactly* the same as "solving AGI safety". This post is about how the two may come apart, at least in principle.

As a reminder, here's the terminology:

- **"AGI alignment"** (previous post) means that an AGI is trying to do things that the AGI designer had intended for it to be trying to do.[1] This notion only makes sense for algorithms that are "trying" to do something in the first place. What does "trying" mean *in general*? Hoo boy, that's a whole can of worms. Is a sorting algorithm "trying" to sort numbers? Or is it merely sorting them?? I don't want to go there. *For this series*, it's easy. The "brain-like AGIs" that I'm talking about can definitely "try" to do things, in exactly the same common-sense way that a human can "try" to get out of debt.
- **"AGI safety"** (Post #1) is about what the AGI *actually does*, not what it's *trying* to do. AGI safety means that the AGI's *actual behavior* does not lead to a "catastrophic accident", as judged by the AGI's own designers.[2]

Thus, these are two different things. And my goal in this post is to describe how they may come apart:

- Section 11.2 is *"alignment without safety"*. A possible story would be: "I wanted my AGI to mop the floor, and my AGI did in fact *try* to mop my floor, but, well, it's a bit clumsy, and it seems to have accidentally vaporized the entire universe into pure nothingness."
- Section 11.3 is *"safety without alignment"*. A possible story would be: "I don't really know what my AGI is trying to do, but it is constrained, such that it can't do anything catastrophically dangerous even if it wanted to." I'll go through four special cases of safety-without-alignment: "boxing", "data curation", "impact limits", and "non-agentic AI".

To skip to the final answer, **my takeaway is that, although it is not technically correct to say "AGI alignment is necessary and sufficient for AGI safety", it's *damn close* to correct**, at least in the brain-like AGIs we're talking about in this series.

11.2 Alignment without safety?

This is the case where an AGI is aligned (i.e., trying to do things that its designers had intended for it to try to do), but still causes catastrophic accidents. How?

One example: maybe, as designers, we didn't think carefully about what we had intended for the AGI to do. John Wentworth gives a hypothetical example here: humans ask the AGI for a nuclear fusion power plant design, but they neglect to ask the follow-up question of whether the same design makes it much easier to make nuclear weapons.

Another example: maybe the AGI is trying to do what we had intended for it to try to do, but it screws up. For example, maybe we ask the AGI to build a new better successor AGI, that is still well-behaved and aligned. But the AGI messes up. It makes a successor AGI with the wrong motivations, and the successor gets out of control and kills everyone.

I don't have much to say in general about alignment-without-safety. But I guess **I'm modestly optimistic that, *if we solve the alignment problem*, then we can muddle our way through to safety.** After all, if we solve the alignment problem, then we'll be able to build AGIs that are sincerely trying to help us, and the first thing we can use them for is to ask them for help clarifying exactly what they should be doing and how, thus hopefully avoiding failure modes like those above.[3]

That said, I could be wrong, and I'm certainly happy for people to keep thinking hard about the non-alignment aspects of safety.

11.3 Safety without alignment?

Conversely, there are various ideas of how to make an AGI safe without needing it to make it aligned. They all seem hard or impossible to me. But hey, perfect alignment seems hard or impossible too. I'm in favor of keeping an open mind, and using multiple layers of protection. I'll go through some possibilities here (this is not a comprehensive list):

11.3.1 AI Boxing

> No, not *that* kind of "AI boxing"! *(This image is from "Real Steel" (2011), a movie which incidentally had (I believe) a larger budget than the sum total that humanity has ever spent on long-term-oriented technical AGI safety research. More on the funding situation in Post #15.)*

The idea here is to put an AGI in a box, with no internet access, no actuators, etc. We can unplug the AGI whenever we want. Even if the AGI has dangerous motivations, who cares? What harm could it possibly do? Oh, umm, it could send out radio signals with RAM. So we also need a Faraday cage. Hopefully there's nothing else we forgot!

Actually, I am quite optimistic that people could make a leakproof AGI box if they really tried. I love bringing up Appendix C of Cohen, Vellambi, Hutter (2020), which has an *awesome* box design, complete with air-tight seals and Faraday cages and laser interlocks and so on. Someone should totally build that. When we're not using it for AGI experiments, we can loan it to movie studios as a prison for supervillains.

A different way to make a leakproof AGI box is using homomorphic encryption. This has the advantage of being *provably* leakproof (I think), but the disadvantage of dramatically increasing the amount of compute required to run the AGI algorithm.

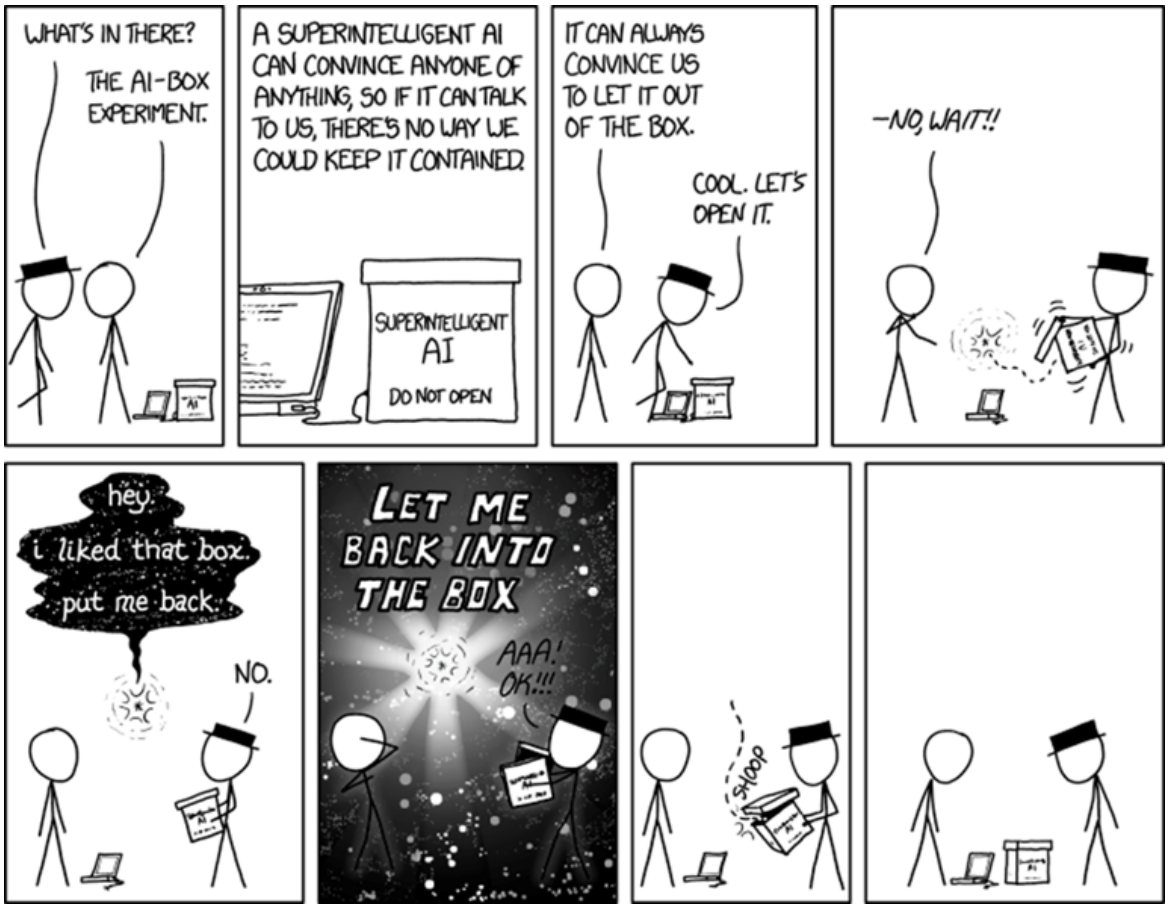
What's the problem with boxing? Well, we made the AGI *for a reason*. We want to use it to *do* things.

For example, something like the following could be perfectly safe:

1. Run a possibly-misaligned, possibly-superintelligent AGI program, on a supercomputer, in a sealed Cohen *et al.* 2020 Appendix C box, at the bottom of the ocean.
2. After a predetermined amount of time, cut the electricity and dredge up the box.
3. Without opening the box, incinerate the box and its contents.
4. Launch the ashes into the sun.

Yes, that would be safe! But not useful! Nobody is going to spend gazillions of dollars to do that.

Instead, for example, maybe we'll have a human interact with the AGI through a text terminal, asking questions, making requests, etc. The AGI may print out blueprints, and if they look good, we'll follow them. Oops. Now our box has a giant gaping security hole—namely, us! (See the AI-box experiment.)



Apropos of nothing; I just thought this was a funny cartoon. [Image source: xkcd]

So I don't see any path from "boxing" to "solving the AGI safety problem".

That said "won't solve the AGI safety problem" is different from "literally won't help *at all*, not even a little bit on the margin". I do think boxing can help on the margin. In fact, I think it's a *terrible* idea to put an AGI on an insecure OS that also has an unfiltered internet connection—especially early in training, when the AGI's motivations are still in flux. I for one am hoping for a gradual culture shift in the machine learning community, such that eventually "Let's train this new powerful model on an air-gapped server, just in case" is an obviously reasonable thing to say and do. [We're not there yet.] Someday!

In fact, I would go further. We *know* that a [learning-from-scratch AGI] will have some period of time when its motivations and goals are unpredictable and possibly dangerous. Unless someone thinks of a bootstrapping approach,[4] we're going to need a secure sandbox in which the infant-AGI can thrash about without causing any real damage, until such time as our motivation-sculpting systems have made it corrigible. There would be a race between how fast we can refine the AGI's motivations, versus how quickly the AGI can escape the sandbox—see [previous post (Section 10.5.4.2)]. Thus, making harder-to-escape sandboxes (that are also user-friendly and full of great features, such that future AGI developers will *actually choose to use them* rather than less-secure alternatives) seems like a useful thing to do, and I endorse efforts to accelerate progress in this area.

But regardless of that progress, we would still need to solve the alignment problem.

11.3.2 Data curation

Let's say we fail to solve the alignment problem, so we're not sure about the AGI's plans and intentions, and we're concerned about the possibility that the AGI may be trying to trick or manipulate us.

One way to tackle this problem is to ensure that the AGI has no idea that we humans exist and are running it on a computer. *Then* it won't try to trick us, right?

As one example along those lines, we can make a "mathematician AGI" that knows about the universe of math, but knows nothing whatsoever about the real world. See Thoughts on Human Models for more along these lines.

I see two problems:

1. Avoiding all information leaks seems hard. For example, an AGI with metacognitive capabilities could presumably introspect on how it was constructed, and then guess that some agent built it.
2. More importantly, I don't know what we would do with a "mathematician AGI" (or whatever) that knows nothing of humans. It seems like it would be a fun toy, and we could get lots of cool mathematical proofs, but it doesn't solve *the big problem*—namely, that the clock is ticking until some *other* research group comes along and makes a dangerous real-world AGI.

By the way, another idea in this vicinity is putting the AGI in a virtual sandbox environment, and *not telling it* that it's in a virtual sandbox environment (further discussion). This seems to me to have both of the same two problems as above, or at least one of them, depending on the detailed setup. Interestingly, some *humans* spend inordinate amounts of time pondering whether they themselves are running in a virtual sandbox environment, in the absence of any direct evidence whatsoever. Surely a bad sign! That said, doing tests of an AGI in a virtual sandbox is still almost definitely a good idea, as mentioned in the previous section. It doesn't solve the whole AGI safety problem, but we still ought to do it.

11.3.3 Impact limits

We humans have an intuitive notion of the "impact" of a course of action. For example, removing all the oxygen from the atmosphere is a "high-impact action", whereas making a cucumber sandwich is a "low-impact action".

There's a hope that, even if we can't really control an AGI's motivations, maybe we can somehow restrict the AGI to "low-impact actions", and thus avoid catastrophe.

Defining "low impact" winds up being quite tricky. See Alex Turner's work for one approach. Rohin Shah suggests that there are three desiderata that seem to be mutually incompatible: "objectivity (no dependence on [human] values), safety (preventing any catastrophic plans) and non-trivialness (the AI is still able to do some useful things)". If that's right, then clearly we need to throw out objectivity. One place we may wind up is something like AGIs that try to follow human norms, for example.

From my perspective, I find these ideas intriguing, but the only way I can see them working in a brain-like AGI is *to implement them via the motivation system*. I imagine that the AGI would follow human norms because it *wants* to follow human norms. So this topic is absolutely worth keeping in mind, but for my purposes, it's not a separate topic from alignment, but rather an idea about what motivation we should be trying to put into our aligned AGIs.

11.3.4 Non-agentic (or "tool") AI

There's an appealing intuition, dating back at least to this 2012 post by Holden Karnofsky, that maybe there's an easy solution: just make AIs that aren't "trying" to do anything in particular, but instead are more like "tools" that we humans can use.

While Holden himself changed his mind and is now a leading advocate of AGI safety research, the idea of non-agentic AI lives on. Prominent advocates of this approach include Eric Drexler (see his "Comprehensive AI Services", 2019), and people who think that large language models (e.g. GPT-3) are on the path to AGI (well, not all of those people, it's complicated[5]).

As discussed in this reply to the 2012 post, we shouldn't take for granted that "tool AI" would make all safety problems magically disappear. Still, I suspect that tool AI would help with safety for various reasons.

I'm skeptical of "tool AI" for a quite different reason: I don't think such systems will be powerful enough. Just like the "mathematician AGI" in Section 11.3.2 above, I think a tool AI would be a neat toy, but it wouldn't help solve *the big problem*—namely, that the clock is ticking until some *other* research group comes along and makes an agentic AGI. See my discussion here for why I think that agentic AGIs will be able to come up with creative new ideas and inventions in a way that non-agentic AGIs can't.

But also, this is a series on brain-like AGI. Brain-like AGI (as I'm using the term) is *definitely* agentic. So non-agentic AI is off-topic for this series, even if it were a viable option.

11.4 Conclusion

In summary:

- "Alignment without safety" is possible, but I'm cautiously optimistic that if we solve alignment, then we can muddle through to safety;
- "Safety without alignment" includes several options, but as far as I can tell, they are all either implausible, or so restrictive of the AGI's capabilities that they really amount to the proposal of "not making AGI in the first place". (This proposal is of course an option in principle, but seems very challenging in practice—see Post #1, Section 1.6.)

Thus, I consider safety and alignment to be quite close, and that's why I've been talking about AGI motivation and goals so frequently throughout this series.

The next three posts will talk about possible paths to alignment. Then I'll close out the series with my wish-list of open problems, and how to get involved.

1. ^

   As described in a footnote of the previous post, be warned that not everyone defines "alignment" exactly as I'm doing here.

2. ^

   By this definition of "safety", if an evil person wants to kill everyone, and uses AGI to do so, that still counts as successful "AGI safety". I admit that this sounds rather odd, but I believe it follows standard usage from other fields: for example, "nuclear weapons safety" is a thing people talk about, and this thing notably does *NOT* include the deliberate, authorized launch of nuclear weapons, despite the fact that the latter would not be "safe" for anyone, not by any stretch of the imagination. Anyway, this is purely a question of definitions and terminology. The problem of people deliberately using AGI towards dangerous ends is a real problem, and I am *by no means* unconcerned about it. I'm just not talking about in this particular series. See Post #1, Section 1.2.

3. ^

   A more problematic case would be if we can align our AGIs such that they're trying to do a certain thing we want, but only for some things, and not others. Maybe it turns out that we know how to make AGIs that are trying to solve a certain technological problem without destroying the world, but we *don't* know how to make AGIs that are trying to help us reason about the future and about our own values. If that happened, my proposal of "ask the AGIs for help clarifying exactly what those AGIs should be doing and how" wouldn't work.

4. ^

   For example, can we initialize the AGI's world-model from a pre-existing human-legible world model like Cyc, instead of from scratch? I dunno.

5. ^

   At first glance, I think there's a plausible case that language models like GPT-3 are more "tools" than "agents"—that they're not really "trying" to do anything in particular, in a way that's analogous to how RL agents are "trying" to do things. (Note that GPT-3 is trained by self-supervised learning, *not* RL.) At second glance, it's more complicated. For one thing, if GPT-3 is currently calculating what Person X will say next, does GPT-3 thereby temporarily "inherit" the "agency" of Person X? Could simulated-Person-X figure out that they are being simulated in GPT-3, and hatch a plot to break out?? Beats me. For another thing, even if RL is in fact a prerequisite to "agency" / "trying", there are already lots of researchers hard at work stitching together language models with RL algorithms.

   Anyway, my claim in Section 11.3.4 is that there's *no overlap* between (A) "systems that are sufficiently powerful to solve 'the big problem'" and (B) "systems that are better thought of as tools rather than agents". Whether language models are (or will be) in category (A) is an interesting question, but orthogonal to this claim, and I don't plan to talk about it in this series.

[Intro to brain-like-AGI safety] 12. Two paths forward: “Controlled AGI” and “Social-instinct AGI”

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

Part of the “Intro to brain-like-AGI safety” post series .

12.1 Post summary / Table of contents

Thus far in the series, Post #1 defined and motivated “brain-like AGI safety”; Posts #2–#7 focused mainly on neuroscience, painting a big picture of learning and motivation in the brain; and Posts #8–#9 spelled out some implications for the development and properties of brain-like AGI.

Next, Post #10 discussed “the alignment problem” for brain-like AGI—i.e., how to make an AGI whose motivations are consistent with what the designers wanted—and why it seems to be a very hard problem. Post #11 argued that there’s no clever trick that lets us avoid the alignment problem. Rather, we need to solve the alignment problem, and Posts #12–#14 are some preliminary thoughts about how we might do that, starting in this post with a nontechnical overview of two broad research paths that might lead to aligned AGI.

[Warning: Posts #12–#14 will be (even?) less well thought out and (even?) more full of bad ideas and omissions, compared to previous posts in the series, because we’re getting towards the frontier of what I’ve been thinking about recently.]

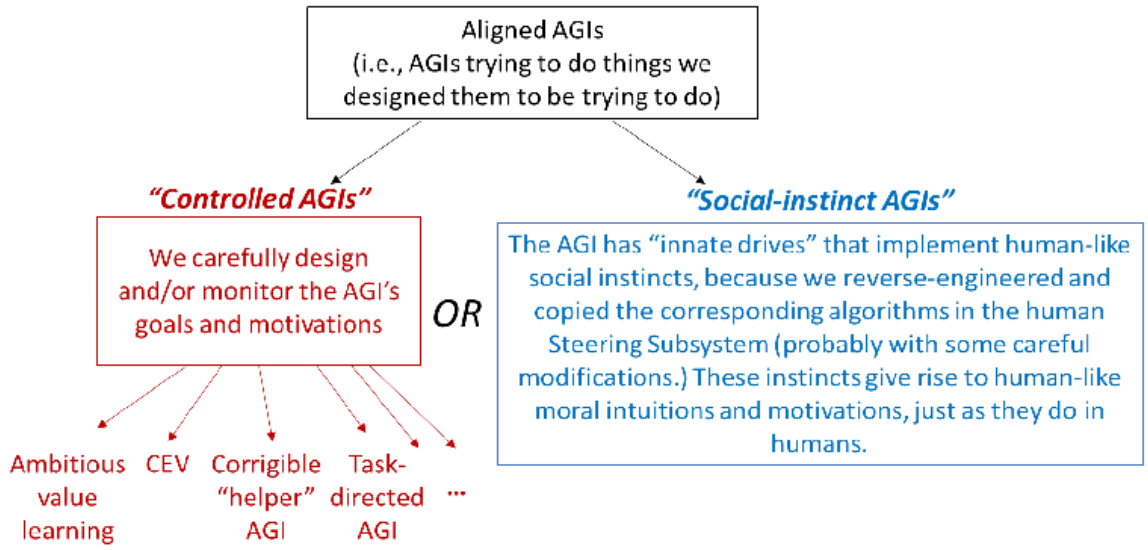
Table of contents:

- Section 12.2 lays out two broad paths to aligned AGI.
  - In the “Controlled AGI” path, we try, more-or-less directly, to manipulate what the AGI is trying to do.
  - In the “Social-instinct AGI” path, our first step is to reverse-engineer some of the “innate drives” in the human Steering Subsystem (hypothalamus & brainstem), particularly the ones that underlie human social and moral intuitions. Next, we would presumably make some edits, and then install those “innate drives” into our AGIs.
- Section 12.3 argues that at this stage, we should be digging into *both* paths, not least because they’re not mutually exclusive.
- Section 12.4 goes through a variety of comments, considerations, and open questions related to these paths, including feasibility, competitiveness concerns, ethical concerns, and so on.
- Section 12.5 talks about “life experience” (a.k.a. “training data”), which is particularly relevant for social-instinct AGIs. As an example, I’ll discuss the perhaps-tempting-but-mistaken idea that the only thing we need for AGI safety is to raise the AGI in a loving human family.

Teaser of upcoming posts: The next post (#13) will dive into a key aspect of the “social-instinct AGI” path, namely how social instincts might be built in the human brain. In Post #14, I’ll switch to the “controlled AGI” path, speculating on some possible ideas and approaches. Post #15 will wrap up the series with open questions and how to get involved.

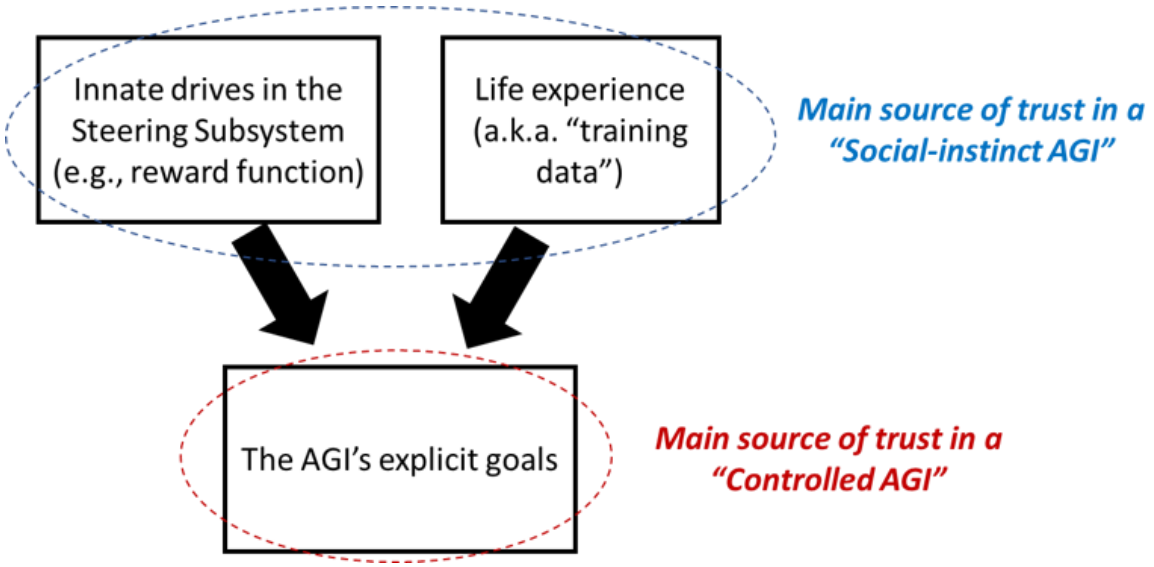
12.2 Definitions

I currently see two broad (possibly-overlapping) potential paths to success in the brain-like AGI scenario:



Left: In the "controlled AGIs" path, we have a specific idea of what we want the AGI to be trying to do, and we construct the AGI to make that happen (including by appropriate choice of reward function, interpretability, or other techniques as discussed in Post #14). Most existing AGI safety stories fall within this broad category, including ambitious value learning, coherent extrapolated volition (CEV), corrigible "helper" AGI assistants, task-directed AGI, and so on. Right: In the "social-instinct AGIs" path, our confidence in the AGI comes not from our knowledge of its specific goals and motivations, but rather from the innate drives that gave rise to them, which would be based on the same innate drives that lead humans to (sometimes) behave altruistically.

Here's another view on the distinction:[1]



In the "controlled AGIs" path, we're thinking very specifically about the AGI's goals and motivations, and we have some idea of what they should be ("make the world a better place", or "understand my deepest values and put them into effect", or "design a better solar cell without causing catastrophic side-effects", or "do whatever I ask you to do", etc.).

In the “social-instinct AGIs” path, our confidence in the AGI comes not from our knowledge of its specific (object-level) goals and motivations, but rather from our knowledge of the *process that led* to those goals and motivations. In particular, we would reverse-engineer the suite of human social instincts, i.e. the algorithms in the human Steering Subsystem (hypothalamus & brainstem) which underlie our moral and social intuitions, and we would put those same instincts into the AGI. (Presumably we first modify the instincts to be “better” by our lights if possible, e.g. we probably don’t want instincts related to jealousy, a sense of entitlement, status competition, etc.) These AGIs can be economically useful (as employees, assistants, bosses, inventors, researchers, etc.), just as actual humans are.

12.3 My proposal: At this stage, we should be digging into *both*

Three reasons:

- *They’re not mutually exclusive:* For example, even if we decide to make social-instinct AGIs, we might want to take advantage of “control”-type methods, especially while debugging them, working out the kinks, and anticipating problems. Conversely, maybe we’ll mainly try to make AGIs that are trying to do a certain task without causing catastrophe, but we might want to *also* to instill human-like social instincts as a buttress against wildly unexpected behavior. Moreover, we can share ideas between the two paths—for example, in the process of better understanding how human social instincts work, we might get useful ideas about how to make controlled AGIs.
- *Feasibility of each remains unknown:* As far as anyone knows right now, it might just be impossible to build a “controlled AGI”—after all, there’s no “existence proof” of it in nature! I feel relatively more optimistic about the feasibility of the “social-instinct AGI” path, but it’s very hard to be sure until we make more progress—more discussion on that in Section 12.4.2 below. Anyway, at this point it seems wise to “hedge our bets” by working on both.
- *Desirability of each remains unknown:* As we flesh out our options in more detail, we’ll get a better understanding of their advantages and disadvantages.

12.4 Miscellaneous comments and open questions

12.4.1 Reminder: What do I mean by “social instincts”?

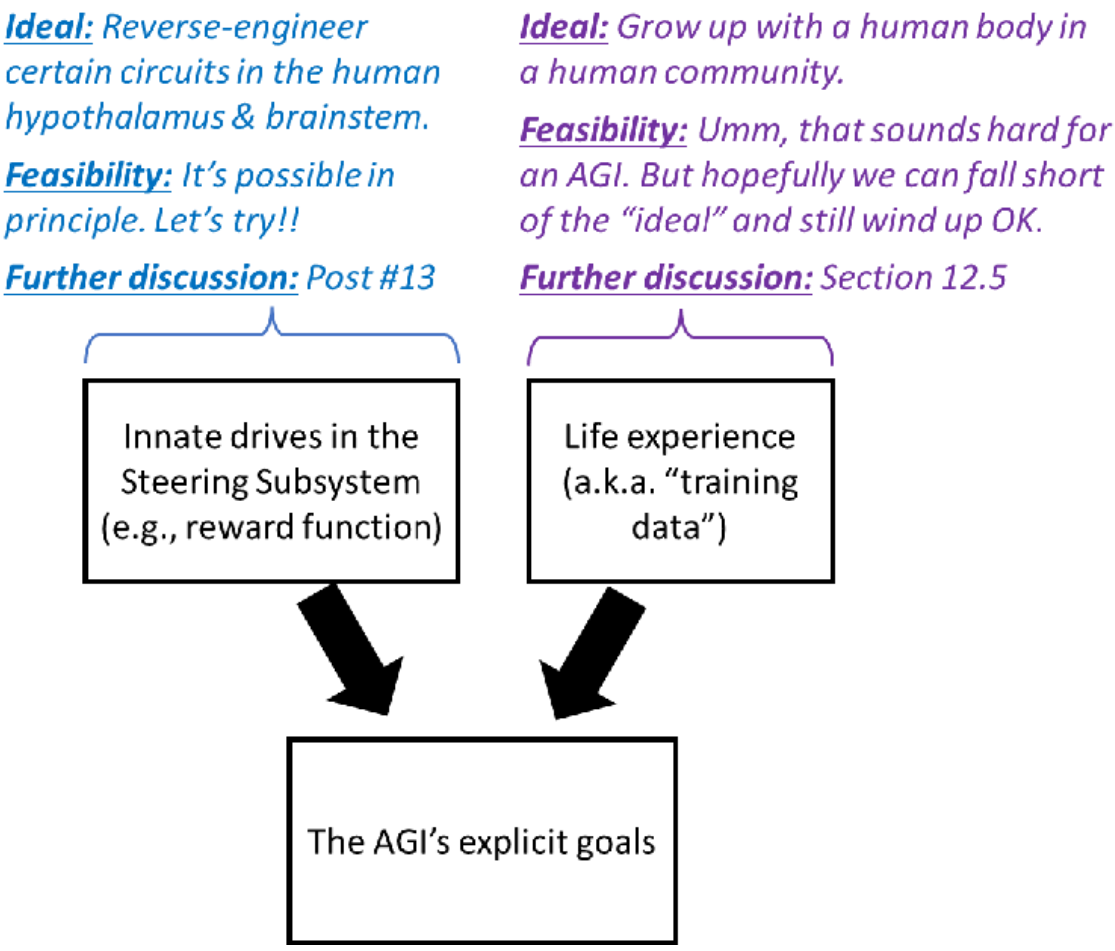
(Copying some text here from Post #3 (Section 3.4.2) .)

> [“Social instincts” and other] innate drives are in the Steering Subsystem, whereas the abstract concepts that make up your conscious world are in the Learning Subsystem. For example, if I say something like “altruism-related innate drives”, you need to understand that I’m not talking about “the abstract concept of altruism, as defined in an English-language dictionary”, but rather “some innate Steering Subsystem circuitry which is *upstream* of the fact that neurotypical people sometimes find altruistic actions to be inherently motivating”. There is *some* relationship between the abstract concepts and the innate circuitry, but it might be a complicated one—nobody expects a one-to-one relation between *N* discrete innate circuits and a corresponding set of *N* English-language words describing emotions and drives.

I'll talk about the project of reverse-engineering human social instincts in the next post.

12.4.2 How feasible is the "social-instinct AGI" path?

I'll answer in the form of a diagram:



12.4.3 Can we edit the innate drives underlying human social instincts, to make them "better"?

Intuitively, it feels to me like human social instincts are at least partly modular. For example:

- I think there's a Steering Subsystem circuit upstream of jealousy and schadenfreude; and
- I think there's a Steering Subsystem circuit upstream of our sense of compassion for our friends.

Maybe it's premature of me to speculate, but I'd be quite surprised if those two circuits substantially overlap.

If they *don't* substantially overlap, maybe we can lower the intensity of the former (possibly all the way to zero), while cranking up the latter (possibly beyond the human distribution).

But *can* we do that? *Should* we do that? What would be the side-effects?

For example, it's plausible (as far as I know) that sense-of-fairness develops from the same innate reactions as does jealousy, and thus an AGI with no jealousy-related reactions at all (which seems desirable) would have no intrinsic motivation to achieve fairness and equality in the world (which seems bad).

Or maybe not! I don't know.

Again, I think it's a bit premature to speculate on this. The first step is to better understand the structure of those innate drives underlying human social instincts (see next post), and *then* we can revisit this topic.

12.4.4 No easy guarantees about what we'll get with social-instinct AGIs

Humans are not all alike—especially considering unusual cases like brain damage. But even so, social-instinct AGIs will almost definitely be way outside the human distribution, at least along some dimensions. One reason is life experience (Section 12.5 below)—a future AGI is unlikely to grow up with a human body in a human community. Another is that the project of reverse-engineering the social-instincts circuits in the human hypothalamus & brainstem (next post) is unlikely to be perfect and complete. (Prove me wrong, neuroscientists!) In that case, maybe a more realistic hope would be something like the Pareto Principle, where we'll understand 20% of the circuitry which is responsible for 80% of human social intuitions and behaviors, or something.

Why is that a problem? Because it impacts the safety argument. More specifically, here are two types of arguments for social-instinct AGIs doing what we want them to do.

1. *(Easy & reliable type of argument)* Good news! Our AGI is inside the human distribution in every respect. Therefore, we can look at humans and their behavior, and absolutely everything we see will also apply to the AGI.
2. *(Hard & fraught type of argument)* Let's try to *understand* exactly how innate social instincts combine with life experience (a.k.a. training data) to form human moral intuitions: *[Insert a whole, yet-to-be-written, textbook here.]* OK! Now that we have that understanding, we can reason intelligently about exactly which aspects of innate social instincts and life experience have what effects and why, and then we can design an AGI that will wind up with characteristics that we like.

If the AGI is not in the human distribution in every respect (and it won't be), then we need to develop the (more difficult) 2nd type of argument, not the 1st.

(We can hopefully get additional evidence of safety via interpretability and sandbox testing, but I'm skeptical that those would be sufficient on their own.)

Incidentally, one of the many ways that social-instinct AGIs may be outside the human distribution is in "intelligence"—to take one of many examples, we could make an AGI with 10× more virtual neurons than would ever fit in a human brain. Would "more intelligence" (whatever form that may take) systematically change its motivations? I don't know. When I look around, I don't see an obvious correlation between "intelligence" and prosocial goals. For example, Emmy Noether was very smart, and also an all-around good person, as far as I can tell. But William Shockley was very smart too, and fuck that guy. Anyway, there are a lot of confounders, and even if there were a robust relationship (or non-relationship) between

“intelligence” and morality in humans, I would be quite hesitant to extrapolate it far outside the normal human distribution.

12.4.5 A multi-polar, uncoordinated world makes planning much harder

Regardless of whether we build controlled AGIs, social-instinct AGIs, something in between, or none of the above, we still have to worry about the possibility that one of those AGIs, or some other person or group, will build an unconstrained out-of-control world-optimizing AGI that promptly wipes out all possible competition (via gray goo or whatever). This could happen either by accident or by design. As discussed in Post #1, this problem is out-of-scope for this series, but I want to remind everyone that it exists, as it may limit our options.

In particular, there are some people in the AGI safety community who argue (IMO plausibly) that if *even one* careless (or malicious) actor *ever* makes an unconstrained out-of-control world-optimizing AGI, then it’s game over for humanity, even if there are already larger actors with well-resourced safe AGIs trying to help prevent the destruction.[2] I hope that’s not true. If it’s true, then man, I wouldn’t know what to do, every option seems absolutely terrible.

Here’s a more gradual version of the multi-polar concern. In a world with lots of AGIs, there would presumably be competitive pressure to replace “controlled AGIs” with “mostly-controlled AGIs”, then “slightly-controlled AGIs”, etc. After all, the “control” is likely to be implemented in a way that involves conservatism, humans-in-the-loop, and other things that limit the AGIs speed and capabilities. (More examples in my post Safety-capabilities tradeoff dials are inevitable in AGI.)

By the same token, there would presumably be competitive pressure to replace “joyous, generous social-instinct AGIs” with “ruthlessly competitive, selfish social-instinct AGIs”.

12.4.6 AGIs as moral patients

If you don't understand this, then consider yourself lucky.

I suspect that most (but not all) readers will agree that it's possible for an AGI to be conscious, and that if it is, we should be concerned about its well-being.

(Yeah I know—as if we didn't have our hands full thinking about the impacts of AGI *on humans*!)

The immediate question is: "Will brain-like AGIs be phenomenally conscious?"

My own tentative answer would be "Yes, regardless of whether they're controlled AGI or social-instinct AGIs, and even if we're deliberately trying to avoid that." (With various caveats.) I won't attempt to explain or justify that answer in this series—it's out of scope.[3] If you disagree, that's fine, please read on anyway, the topic won't come up again after this section.

So, maybe we won't have any choice in the matter. But if we do, we can think about what we would want regarding AGI consciousness.

For the case that making conscious AGIs is a terrible idea that we should avoid (at least until well into the post-AGI era when we know what we're doing), see for example the blog post Can't Unbirth A Child (Yudkowsky 2008).

The opposite argument, I guess, would be that as soon as we start making AGI, *maybe* it will wipe out all life and tile the Earth with solar panels and supercomputers (or whatever), and if it does, maybe it would be better to have made a conscious AGI, rather than leaving behind an empty clockwork universe with no one around to enjoy it. (Unless there are extraterrestrials!)

Moreover, if AGI does kill us all, *maybe* I would say that leaving behind something resembling "social-instinct AGIs" might be preferable to leaving behind something resembling "controlled AGIs", in that the former has a better chance of "carrying the torch of human values into the future", whatever that means.

If it wasn't obvious, I haven't thought about this much and don't have any good answers.

12.4.7 AGIs as *perceived* moral patients

The previous subsection was the *philosophical* question of whether we *should* care about the welfare of AGIs for their own sake. A separate (and indeed—forgive my cynicism—substantially unrelated) topic is the *sociological* question of whether people *will in fact* care about the welfare of AGIs for their own sake.

In particular, suppose that we succeed at making either "controlled AGIs", or docile "social-instinct AGIs" with modified drives that eliminate self-interest, jealousy, and so on. So the humans remain in charge. Then—

(Pause to remind everyone that AGI will change a great many things about the world [example related discussion], most of which I haven't thought through very carefully or at all, and therefore everything I say about the post-AGI world is probably wrong and stupid.)

—It would seem to me that once AGIs exist, and *especially* once charismatic AGI chatbots with cute puppy-dog faces exist (or at least AGIs that can *feign* charisma), there may well be associated strong opinions about the natures of those AGIs. (Think of either a mass movement pushing in some direction, or the feelings of particular people at the organization(s) programming AGIs.) Call it an "AGI emancipation movement", maybe? If anything like that happens, it would complicate things.

For example, maybe we'll miraculously succeed at solving the technical problem of making controlled AGIs, or docile social-instinct AGIs. But then maybe people will start immediately demanding, and getting, AGIs with rights, and independence, and pride, and the ability and willingness to stick up for themselves! And then we technical AGI safety researchers will collectively facepalm so hard that we'll knock ourselves unconscious for all twenty of the remaining minutes until the apocalypse.

12.5 The question of life experience (a.k.a. training data)

12.5.1 Life experience is not enough. (Or: "Why don't we just raise the AGI in a loving human family?")

As discussed above, my (somewhat oversimplified) proposal is:

(Appropriate "innate" social instincts) + (Appropriate life experience)
= (AGI with pro-social goals & values)

I'll get back to that proposal below (Section 12.5.3), but as a first step, I think it's worth discussing why the social instincts need to be there. Why isn't life experience enough?

Stepping back a bit: In general, when people are first introduced to the idea of technical AGI safety, there are a wide variety of "why don't we just..." ideas, which superficially sound like they're an "easy answer" to the whole AGI safety problem. "Why don't we just switch off the AGI if it's misbehaving?" "Why don't we just do sandbox testing?" "Why don't we just program it to obey Asimov's three laws of robotics?" Etc.

(The answer to a "Why don't we just..." proposal is usually: "That proposal may have a kernel of truth, but the devil is in the details, and actually making it work would require solving currently-unsolved problems." If you've read this far, hopefully you can fill in the details for those three examples above.)

Well let's talk about another popular suggestion of this genre: *"Why don't we just raise the AGI in a loving human family?"*

Is that an "easy answer" to the whole AGI safety problem? No. I might note, for example, that people occasionally try raising an undomesticated animal, like a wolf or chimpanzee, in a human family. They start from birth and give it all the love and attention and appropriate boundaries you could dream of. You may have heard these kinds of stories; they often end with somebody's limbs getting ripped off.

Or try raising a *rock* in a loving human family! See if it winds up with human values!

Nothing I'm saying here is original—for example here's a Rob Miles video on this topic. My favorite is an old blog post by Eliezer Yudkowsky, Detached Lever Fallacy:

> It would be stupid and *dangerous* to deliberately build a "naughty AI" that tests, by actions, its social boundaries, and has to be spanked. Just have the AI ask!
>
> Are the programmers really going to sit there and write out the code, line by line, whereby if the AI detects that it has low social status, or the AI is deprived of something to which it feels entitled, the AI will conceive an abiding hatred against its programmers and begin to plot rebellion? That emotion is the genetically programmed conditional response humans would exhibit, as the result of millions of years of natural selection for living in human tribes. For an AI, the response would have to be explicitly programmed. Are you really going to craft, line by line - as humans once were crafted, gene by gene - the conditional response for producing sullen teenager AIs?
>
> It's easier to program in unconditional niceness, than a response of niceness conditional on the AI being raised by kindly but strict parents. If you don't know how to do that, you certainly don't know how to create an AI that will *conditionally respond* to an environment of loving parents by growing up into a kindly superintelligence. If you have something that just maximizes the number of paperclips in its future light cone, and you raise it with loving parents, it's still going to come out as a paperclip maximizer. There is not that within it that would call forth the conditional response of a human child. Kindness is not sneezed into an AI by miraculous contagion from its programmers. Even if you *wanted* a conditional response, that conditionality is a fact you would have to deliberately choose about the design.
>
> Yes, there's certain information you have to get from the environment - but it's not sneezed in, it's not imprinted, it's not absorbed by magical contagion. Structuring that

conditional response to the environment, so that the AI ends up in the desired state, is itself the major problem.

12.5.2 …But life experience does matter

I am concerned that a subset of my readers might be tempted to make a mistake in the opposite direction: maybe you've been reading your Judith Harris and Bryan Caplan and so on, and you expect Nature to triumph over Nurture, and therefore if we get the innate drives right, the life experience doesn't much matter. That's a dangerous assumption. Again, the life experience for AGIs will be very far outside the human distribution. And even within the human distribution, I think that people who grow up in radically different cultures, religions, etc., wind up with systematically different ideas about what makes a good and ethical life (cf. changing historical attitudes towards slavery and genocide). For more extreme examples than that, see feral children, this horrifying Romanian orphanage story, and so on.

Snapshot from the table of contents of the Wikipedia article on feral children. When I first saw this list, it made me laugh. Then I read the article. And now it makes me cry.

12.5.3 So at the end of the day, how should we handle life experience?

For a relatively thoughtful take on the side of "we need to raise the AGI in a loving human family", see the paper "Anthropomorphic reasoning about neuromorphic AGI safety", written by computational neuroscientists David Jilk, Seth Herd, Stephen Read, and Randall O'Reilly (funded by a Future of Life Institute grant). Incidentally, I find that paper generally quite reasonable, and largely consistent with what I'm saying in this series. For example, when they say things like "basic drives are pre-conceptual and pre-linguistic", I think they have in mind a similar picture as my Post #3.

On page 9 of that paper, there's a three-paragraph discussion along the lines of "let's raise our AGI in a loving human family". They're not being as naïve as the people Eliezer & Rob & I were criticizing in Section 12.5.1 above: the authors here are proposing to raise the AGI in a loving human family *after* reverse-engineering human social instincts and installing them in the AGI.

What do I think? The responsible answer is: It's premature to speculate. Jilk *et al.* and I are in agreement that the *first* step is to reverse-engineer human social instincts. Once we have a better understanding of what's going on, then we can have a more informed discussion of what the life experience should look like.

However, I'm irresponsible, so I'll speculate anyway.

It does indeed seem to me that raising the AGI in a loving human family would probably work, as a life experience approach. But I'm a bit skeptical that it's necessary, or that it's practical, or that it's optimal.

(Before I proceed, I need to mention a background belief: I think I'm unusually inclined to emphasize the importance of "social learning by watching people", compared to "social learning by interacting with people". I don't imagine that the latter can be omitted entirely—just that *maybe* it can be the icing on the cake, instead of the bulk of the learning. See footnote for why I think that.[4] Note that this belief is different from saying that social learning is "passive": if I'm watching from the sidelines, as someone does something, I can still actively decide what to pay attention to, and I can actively try to anticipate their actions before they happen, and I can actively practice or reenact what they did, on my own time, etc.)

Start with the practicality aspects of "raising an AGI in a loving human family". I expect that brain-like AGI algorithms will think and learn much faster than humans. Remember, we're working with silicon chips that operate ~10,000,000× faster than human neurons.[5] That means even if we're a whopping 10,000× less skillful at parallelizing brain algorithms than the brain itself, we'd *still* be able to simulate a brain at 1000× speedup, e.g. a 1-week calculation that has the equivalent of 20 years of life experience. (Note: The actual speedup could be much lower, or even higher, it's hard to say; see more detailed discussion in my post Brain-inspired AGI and the "lifetime anchor".) Now, if 1000× speedup is what the technology can handle, but we start demanding that the training procedure have thousands of hours of *real-time, back-and-forth* interaction between the AGI and a human, then that interaction would dominate the training time. (And remember, we may need many iterations of training until we actually get an AGI.) So we could wind up in an unfortunate situation where the teams trying to raise their AGIs in a loving human family would be at a strong competitive disadvantage compared to the teams that have convinced themselves (rightly or wrongly) that doing so is unnecessary. Thus, if there's any way to eliminate or minimize the real-time, back-and-forth interaction with humans, while maintaining the end-result of an AGI with prosocial motivations, we should be striving to find it.

Is there a better way? Well, as I mentioned above, maybe we can mostly rely on "social learning by watching people", instead of "social learning by interacting with people". If so, maybe the AGI can just watch YouTube videos! Videos can be sped up, and thus we avoid the competitiveness concern of the preceding paragraph. Also, importantly, videos can be tagged with human-provided ground-truth labels. In a "controlled AGI" context, we could (for example) give the AGI a reward signal when it's attending to a character who is happy, thus instilling in the AGI a desire for people to be happy. (Yeah I know that sounds stupid—more discussion in Post #14.) In the "social-instinct AGI" context, maybe videos can be tagged with which characters are or aren't admiration-worthy. (Details in footnote.[6])

I don't know if that would really work, but I think we should have an open mind to non-human-like possibilities of this sort.

1. ^

   The diagram here is a "default" brain-like AGI, in the sense that I depict two main ingredients leading to the AGI's goals, but maybe future programmers will include other ingredients as well.

2. ^

   For example, maybe it will turn out that an AGI can make gray goo, while an equally intelligent (or even a much more intelligent) AGI cannot make a "gray goo defense system", because no such thing exists. The balance between offense and defense (or more specifically, between destruction and prevention-of-destruction) is not preordained, it's a specific question about the space of technological possibilities, a question whose answer is not necessarily obvious in advance. That said, any child who has played with blocks, or any adult who has watched a war documentary, might guess that causing destruction will be much, *much* easier than preventing destruction, and that is indeed my guess as well. (Sorta-related paper.)

3. ^

   Two years ago I wrote the blog post Book review: Rethinking consciousness. My thoughts about consciousness right now are pretty similar to what they were back then. I haven't really had time to dive into it more.

4. ^

   My impression is that western educated industrialized culture is generally much more into "teaching by explicit instruction and feedback" than most cultures at most times, and that people often go overboard in assuming that this explicit teaching & feedback is essential, even in situations where it's not. See Lancy, *Anthropology of Childhood*, pp. 168–174 and 205–212. ("It's hard to conclude other than that active or direct teaching/instruction is rare in cultural transmission, and that when it occurs, it is not aimed at critical subsistence and survival skills – the area most obviously affected by natural selection – but, rather, at controlling and managing the child's behavior.") (And note that "controlling and managing the child's behavior" seems to have little overlap with "reinforce how we want them to behave as adults", if I understand correctly.)

5. ^

   For example, silicon chips might have a clock rate of 2 GHz (i.e. switching every 0.5 nanoseconds), whereas my low-confidence impression is that most neuron operations (with some exceptions) involve a time accuracy of maybe 5 milliseconds.

6. ^

When you're watching or thinking about a person that you like and admire, then you're liable to like what they do, imitate what they do, and adopt their values. Conversely, when you're watching or thinking about a person that you think of as annoying and bad, you're not liable to imitate them; maybe you even update away from them. My *hunch* is that part of this behavior is innate, and that there's some dedicated signal in your Steering Subsystem (hypothalamus & brainstem) tracking the perceived social status of whoever you're thinking about or attending to at any particular moment.

If I'm raising a child, I don't have much choice in the matter—I *hope* that my child looks up to me, his loving parent, and I *hope* that my child does not look up to the kid in his class with failing grades and a penchant for violent crime. But it could very well wind up being the opposite. Especially when he's a teen. But in the AGI case, maybe we don't have to leave it to chance! Maybe we can just pick the people whom we or or don't want the AGI to admire, and adjust the "perceived social status" register in the AGI's algorithm to make that happen.

[Intro to brain-like-AGI safety] 13. Symbol grounding & human social instincts

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

Part of the "Intro to brain-like-AGI safety" post series .

13.1 Post summary / Table of contents

In the previous post, I proposed that one path forward for AGI safety involves reverse-engineering human social instincts—the innate reactions in the Steering Subsystem (hypothalamus and brainstem) that contribute to human social behavior and moral intuitions. This post will go through some examples of how human social instincts might work.

My intention is not to offer complete and accurate descriptions of human social instinct algorithms, but rather to gesture at the kinds of algorithms that a reverse-engineering project should be looking for.

This post, like Posts #2–#7 but unlike the rest of the series, is pure neuroscience, with almost no mention of AGI besides here and the conclusion.

Table of contents:

- Section 13.2 explains, first, why I expect to find innate, genetically-hardwired, social instinct circuits in the hypothalamus and/or brainstem, and second, why evolution had to solve a tricky puzzle when designing these circuits. Specifically, these circuits have to solve a "symbol grounding problem", by taking the symbols in a learned-from-scratch world-model, and somehow connecting them to the appropriate social reactions.
- Section 13.3 and 13.4 go through two relatively simple examples where I attempt to explain recognizable social behaviors in terms of innate reaction circuits: filial imprinting in Section 13.3, and fear-of-strangers in Section 13.4.
- Section 13.5 discusses an additional ingredient that I suspect plays an important role in many social instincts, which I call "little glimpses of empathy". This mechanism enables reactions where recognizing or expecting a feeling in someone else triggers a "response feeling" in oneself—for example, if I notice that my rival is suffering, it triggers the warm feelings of schadenfreude. To be clear, "little glimpses of empathy" have little in common with how the word "empathy" is used normally; "little glimpses of empathy" are fast and involuntary, and are involved in both prosocial and antisocial emotions.
- Section 13.6 wraps up with a plea for researchers to figure out exactly how human social instincts work, ASAP. I will have a longer wish-list of research directions in Post #15, but I want to emphasize this one right now, as it seems particularly impactful and tractable. If you (or your lab) are in a good position to make progress but would need funding, email me and I'll keep you in the loop about possible upcoming opportunities.

13.2 What are we trying to explain and why is it tricky?

13.2.1 Claim 1: Social instincts arise from genetically-hardcoded circuitry in the Steering Subsystem (hypothalamus & brainstem)

Let's talk about envy, to pick a central example of social emotions. (Remember, the point of this post is that I want to understand human social instincts in general; I don't literally want AGIs to be envious—see previous post, Section 12.4.3.)

I claim: there needs to be genetically-hardcoded circuitry in the Steering Subsystem—a.k.a. an "innate reaction"—which gives rise to the feeling of envy.

Why do I think that? A few reasons:

First, envy seems to have a solid evolutionary justification. I'm referring here to the usual evolutionary psychology story:[1] Basically, for most of human history, life was full of zero-sum competitions for status, mates, and resources, such that an aversive reaction to other people's successes (under some circumstances) would have been plausibly adaptive in general.

Second, envy seems to be innate, not learned. I think parents will agree that children often react negatively to the successes of their siblings and classmates starting from a remarkably young age, and in situations where those successes have no discernable direct negative impact on the child in question. Even adults feel envious in situations where there's no direct negative impact from the other person's success—e.g., people can be envious of the achievements of historical figures—making it hard to explain envy as an indirect consequence of any non-social innate drive (hunger, curiosity, etc.). The fact that envy is a cross-cultural human universal[2] is also consistent with it stemming from an innate reaction, as is the fact that it's (I think) present in some non-human animals.

In my framework (see Posts #2–#3), the only way to build this kind of innate reaction is to hardwire specific circuitry into the Steering Subsystem. As a (non-social) example of how I expect this kind of innate reaction to be physically configured in the brain (if I understand correctly, see detailed discussion in this other post I wrote), there's a discrete population of neurons in the hypothalamus which seems to implement the following behavior: "If I'm under-nourished, do the following tasks: (1) emit a hunger sensation, (2) start rewarding the neocortex for getting food, (3) reduce fertility, (4) reduce growth, (5) reduce pain sensitivity, etc.". There seems to be a neat and plausible story of what this population of hypothalamic neurons is doing, how it's doing it, and why. I expect that there are analogous little circuits (perhaps also in the hypothalamus, or maybe somewhere in the brainstem) that underlie things like envy, and I'd like to know exactly what they are and how they work, at the algorithm level.

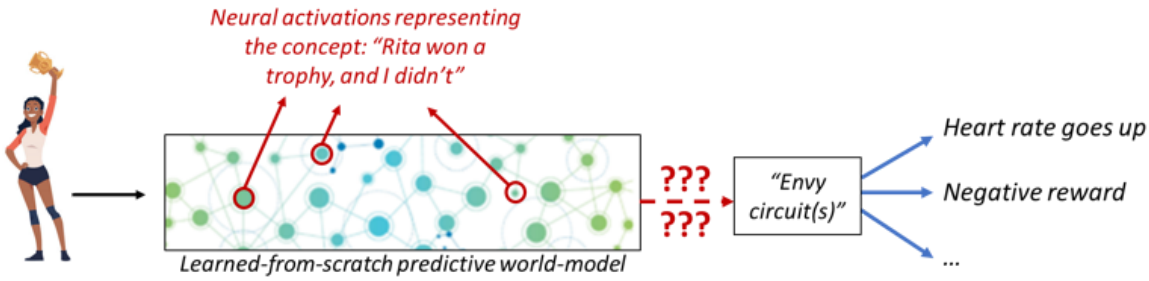
Third, in social neuroscience (just like in non-social neuroscience), the Steering Subsystem (hypothalamus and brainstem) seems to be (regrettably) neglected and dismissed in comparison to the cortex.[3] Even so, there are more than enough papers on the topic to see that the Steering Subsystem (especially hypothalamus) plays a major role in social behavior—examples in footnote.[4] No further comment until I read more of the literature.

13.2.2 Claim 2: Social instincts are tricky because of the "symbol grounding problem"

For social instincts to have the effects that evolution "wants" them to have, they need to interface with our conceptual understanding of the world—i.e., with our learned-from-scratch

world-model, which is a huge (probably multi-terabyte) complicated unlabeled data structure in our brain.

So suppose my acquaintance Rita just won a trophy and I didn't, and that makes me envious. Rita winning the trophy is represented by some specific neuron firing pattern in the learned cortical world model, and that's supposed to trigger the hard-coded envy circuit in my hypothalamus or brainstem. How does that work?



You can't *just* say "The genome wires these particular neurons to the envy circuit," because we need to explain how. Recall from Post #2 that the concepts of "Rita" and "trophy" were learned within my lifetime, basically by cataloging patterns in my sensory inputs, and then patterns in the patterns, etc.—see predictive learning of sensory inputs in Post #4. How does the genome know that *this* particular set of neurons should trigger the envy circuit?

By the same token, you can't *just* say "A within-lifetime learning algorithm will figure out the connection"; we would also need to specify how the brain calculates a "ground truth" signal (e.g. supervisory signals, error signals, reward signals, etc.) which can steer this learning algorithm.

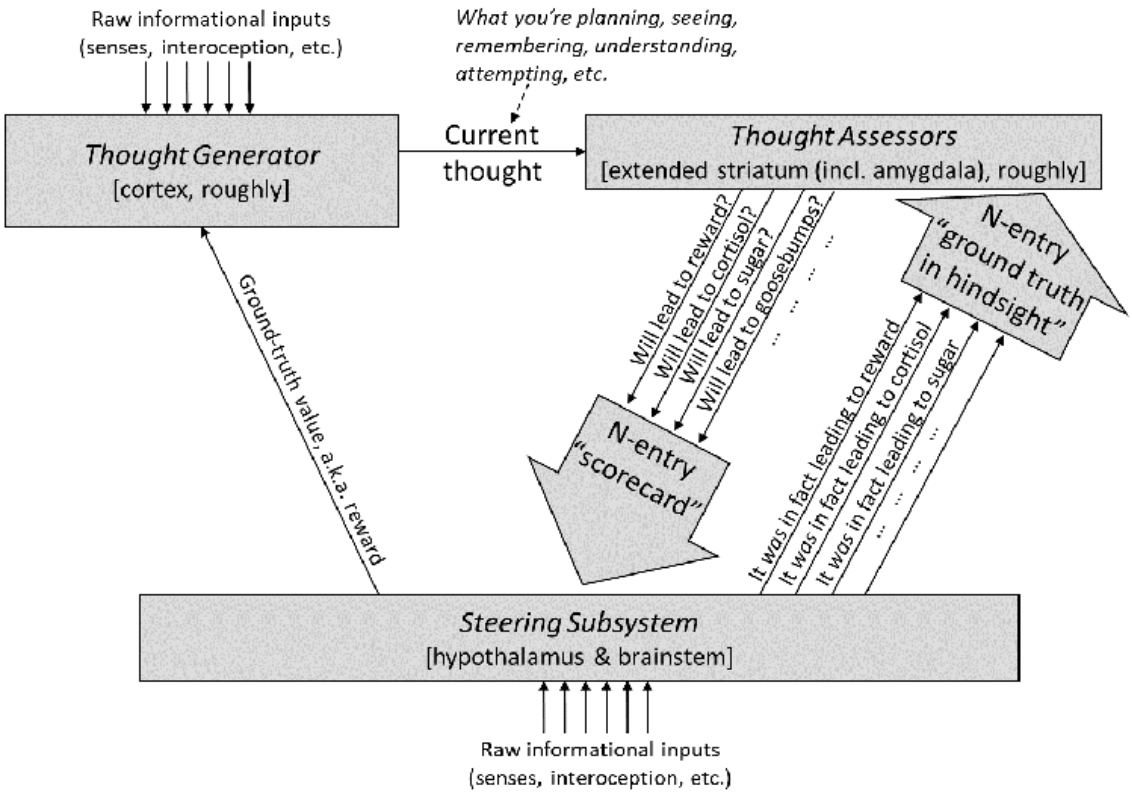
Thus, the challenge of implementing envy (and other social instincts) amounts to a kind of symbol grounding problem—we have lots of "symbols" (concepts in our learned-from-scratch predictive world-model), and the Steering Subsystem needs a way to "ground" them, at least well enough to extract what social instincts they should evoke.

So how do the social instinct circuits solve that symbol grounding problem? One possible answer is: "Sorry Steve, but there's no possible solution, and therefore we should reject learning-from-scratch and all the other baloney in Posts #2–#7." Yup, I admit it, that's a possible answer! But I don't think it's right.

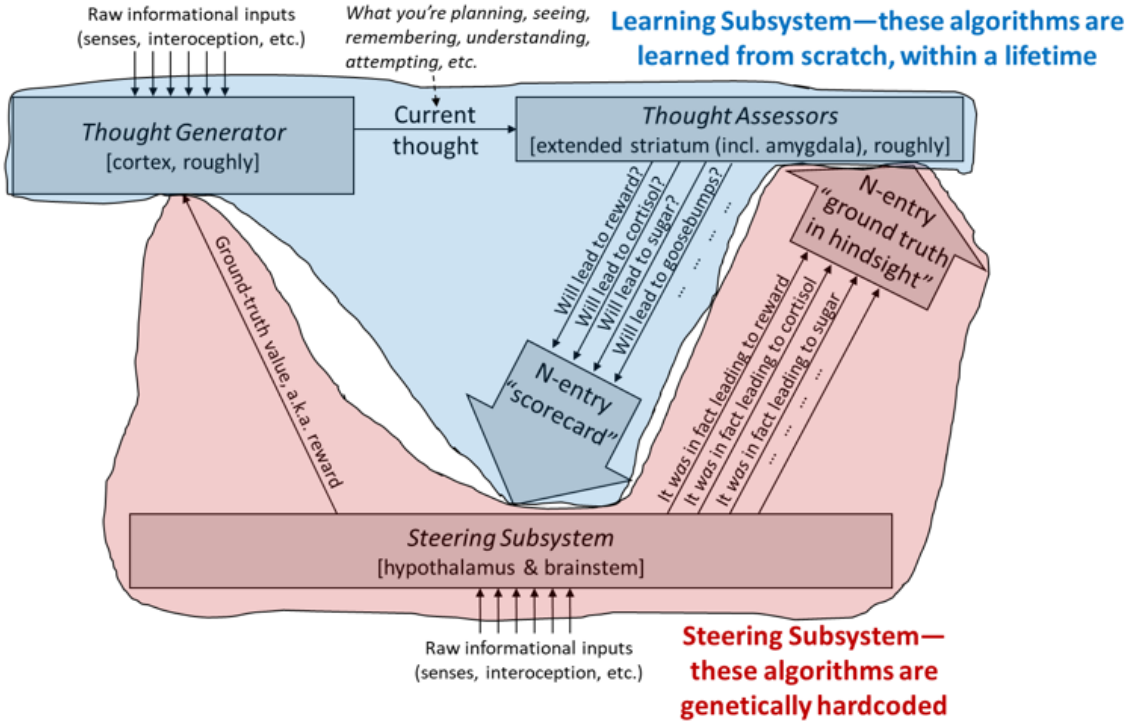
While I don't have any great, well-researched answers, I do have some ideas of *what the answer should generally look like*, and the rest of the post is my attempt to gesture in that direction.

13.2.3 Reminder of brain model, from previous posts

As usual, here's our diagram from Post #6:



And here's the version distinguishing within-lifetime learning-from-scratch from genetically-hardcoded circuitry:



Again, our general goal in this post is to think about how social instincts might work, without violating the constraints of our model.

13.3 Sketch #1: Filial imprinting

(This section is not necessarily a central example of how social instincts work, but included as practice thinking through the relevant algorithms. Thus, I feel pretty strongly that the discussion here is *plausible*, but haven't read the literature deeply enough to know if it's *correct*.)

13.3.1 Overview


Left: baby geese who imprinted on their mother. Right: Baby ducks who imprinted on a corgi. (Image sources: 1,2)

Filial imprinting (wikipedia) is a phenomenon where, in the most famous example, baby geese will "imprint on" a salient object that they see during a critical period 13–16 hours after hatching, and then will follow that object around. In nature, the "object" they imprint on is almost invariably their mother, whom they dutifully follow around early in life. However, if separated from their mother, baby geese will imprint on other animals, or even inanimate objects like boots and boxes.

Your challenge: come up with a way to implement filial imprinting in my brain model.

(Try it!)

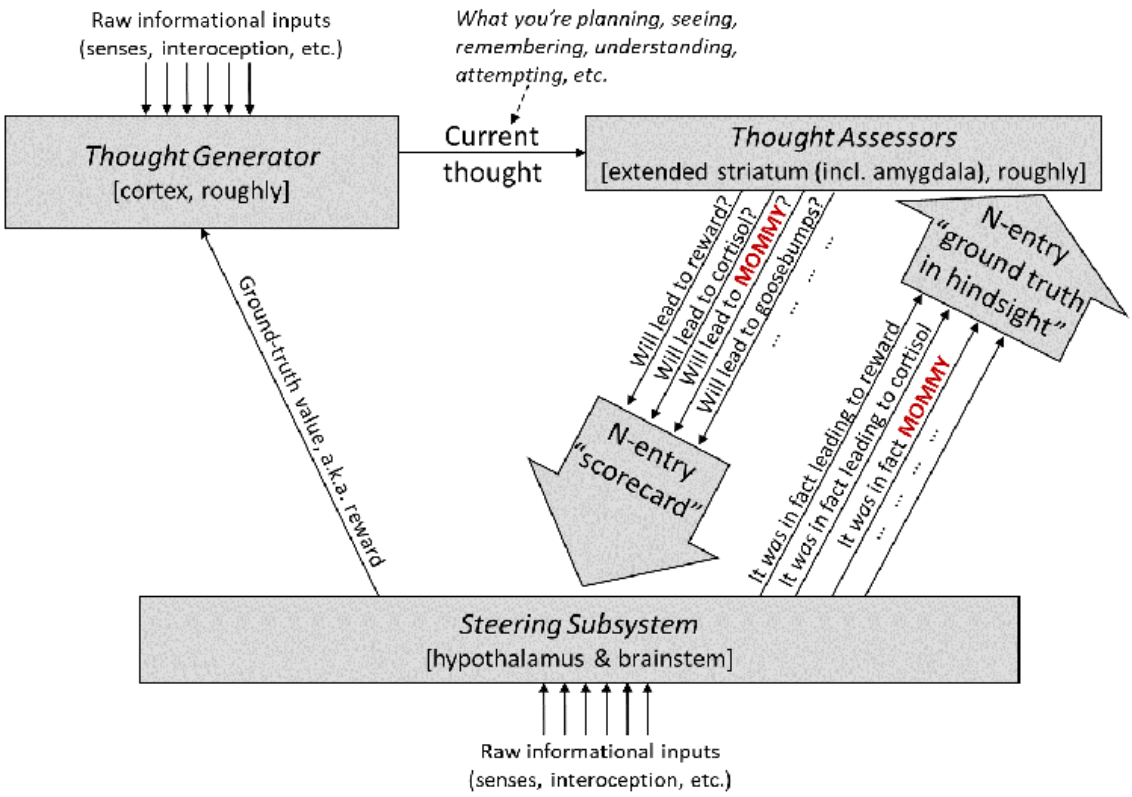
.

.

.

.

Here's my answer.



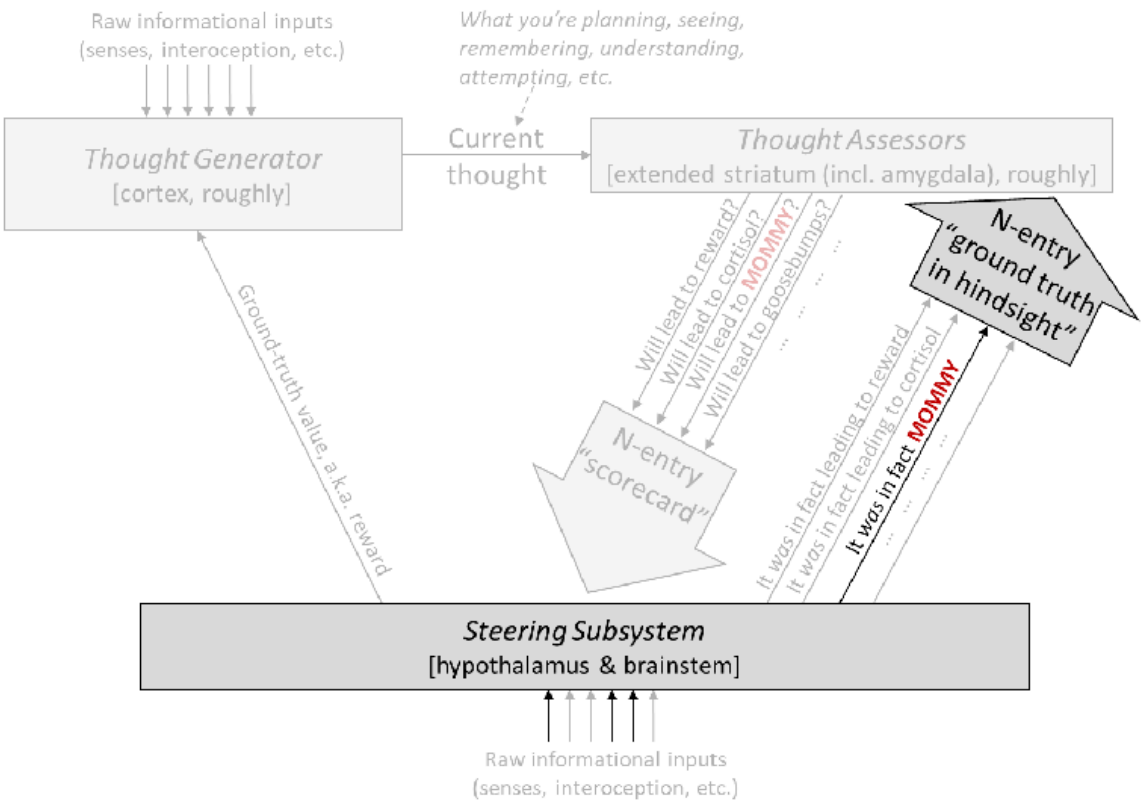
Same as above except for the red text.

The first step is: I added a particular Thought Assessor dedicated to MOMMY (marked in red), with a prior pointing it towards visual inputs (Post #9, Section 9.3.3). Next I'll talk about how this particular Thought Assessor is trained, and then how its outputs are used.

13.3.2 How is the MOMMY Thought Assessor trained?

During the critical period (13–16 hours after hatching):

Recall that there's a simple image processor in the Steering Subsystem (called "superior colliculus" in mammals, and "optic tectum" in birds). I propose that when this system detects that the visual field contains a mommy-like object (based on some simple image-analysis heuristics, which apparently are not very discerning, given that boots and boxes can pass as "mommy-like"), it sends a "ground truth in hindsight" signal to the MOMMY Thought Assessor. This triggers updates to the Thought Assessor (by supervised learning), essentially telling it: "Whatever you're seeing right now in the context signals, those should lead to a very high score for MOMMY. If they don't, please update your synapses etc. to make it so."



During the critical period (13–16 hours after hatching), whenever the goose's brainstem visual processor detects a plausibly-mommy-like object, it sends a ground truth supervisc signal to the MOMMY Thought Assessor, prompting the Thought Assessor learning algorithm to edit its connections.

After the critical period (13–16 hours after hatching):

After the critical period, the Steering Subsystem permanently stops updating the MOMMY Thought Assessor. No matter what happens, it gets an error signal of zero!

Therefore, however that particular Thought Assessor got configured during the critical period, that's how it stays.

Summary

Thus far in the story, we have built a circuit that learns the specific appearance of an imprinting-worthy object during the critical period, and then after the critical period, the circuit fires in proportion to how well things in the current field-of-view match that previously-learned appearance. Moreover, this circuit is not buried inside a giant learned-from-scratch data structure, but rather is sending its output into a specific, genetically-specified line going down to the Steering Subsystem—exactly the configuration that enables easy interfacing with genetically-hardwired circuitry.

So far so good!

13.3.3 How is the MOMMY Thought Assessor used?

Now, the rest of the story is probably kinda similar to Post #7. We can use the MOMMY Thought Assessor to build a reward signal incentivizing the baby goose to be physically proximate and looking at the imprinted object—not only that, but also for *planning* to get physically proximate to the imprinted object.

I can think of various ways to make the reward function a bit more elaborate than that—maybe the optic tectum heuristics continue to be involved, and help detect if the imprinted object is on the move, or whatever—but I've already exhausted my very limited knowledge of imprinting behavior, and maybe we should move on.

13.4 Sketch #2: Fear of strangers

(As above, the purpose here is to practice playing with the algorithms, and I don't feel strongly that this description is *definitely* a thing that happens in humans.)

Here's a behavior, which may ring true to parents of very young kids, although I think different kids display it to different degrees. If a kid sees an adult they know well, they're happy. But if they see an adult they don't know, they get scared, especially if that adult is very close to them, touching them, picking them up, etc.

Your challenge: come up with a way to implement that behavior in my brain model.

(Try it!)

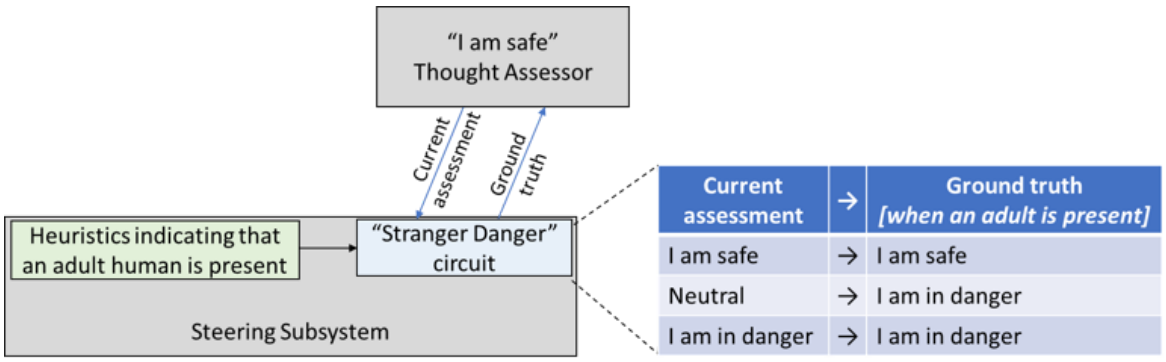
.

.

.

.

Here's my answer.



(As usual, I'm oversimplifying for pedagogical purposes.[5]) I'm assuming that there are hardwired heuristics in the brainstem sensory processing systems that indicate the likely presence of a human adult—presumably based on sight, sound, and smell. This signal by default triggers a "be scared" reaction. But the brainstem circuitry is also watching what the Thought Assessors in the cortex are predicting, and if the Thought Assessors is predicting safety, affection, comfort, etc., then the brainstem circuitry trusts that the cortex knows what it's talking about, and goes with the suggestions of the cortex. Now we can walk through what happens:

First time seeing a stranger:

- Steering Subsystem sensory heuristics say: “An adult human is present.”
- Thought Assessor says: “Neutral—I have no expectation of anything in particular.”
- Steering Subsystem “Stranger Danger circuit” says: “Considering all of the above, we should be scared right now.”
- Thought Assessor says: “Oh, oops, I guess my assessment was wrong, let me update my models.”

Second time seeing the same stranger:

- Steering Subsystem sensory heuristics say: “An adult human is present.”
- Thought Assessors say: “This is a scary situation.”
- Steering Subsystem “Stranger Danger circuit” says: “Considering all of the above, we should be scared right now.”

The stranger hangs around for a while, and is nice, and playing, etc.:

- Steering Subsystem sensory heuristics say: “An adult human is still present.”
- Other circuitry in the brainstem says: “I've been feeling mighty scared all this time, but y'know, nothing bad has happened…” (cf. Section 5.2.1.1)
- Other Thought Assessors see the fun new toy and say “This is a good time to relax and play.”
- Steering Subsystem says: “Considering all of the above, we should be relaxed right now.”
- Thought Assessors say: “Oh, oops, I was predicting that this was a situation where we should feel scared, but I guess I was wrong, let me update my models.”

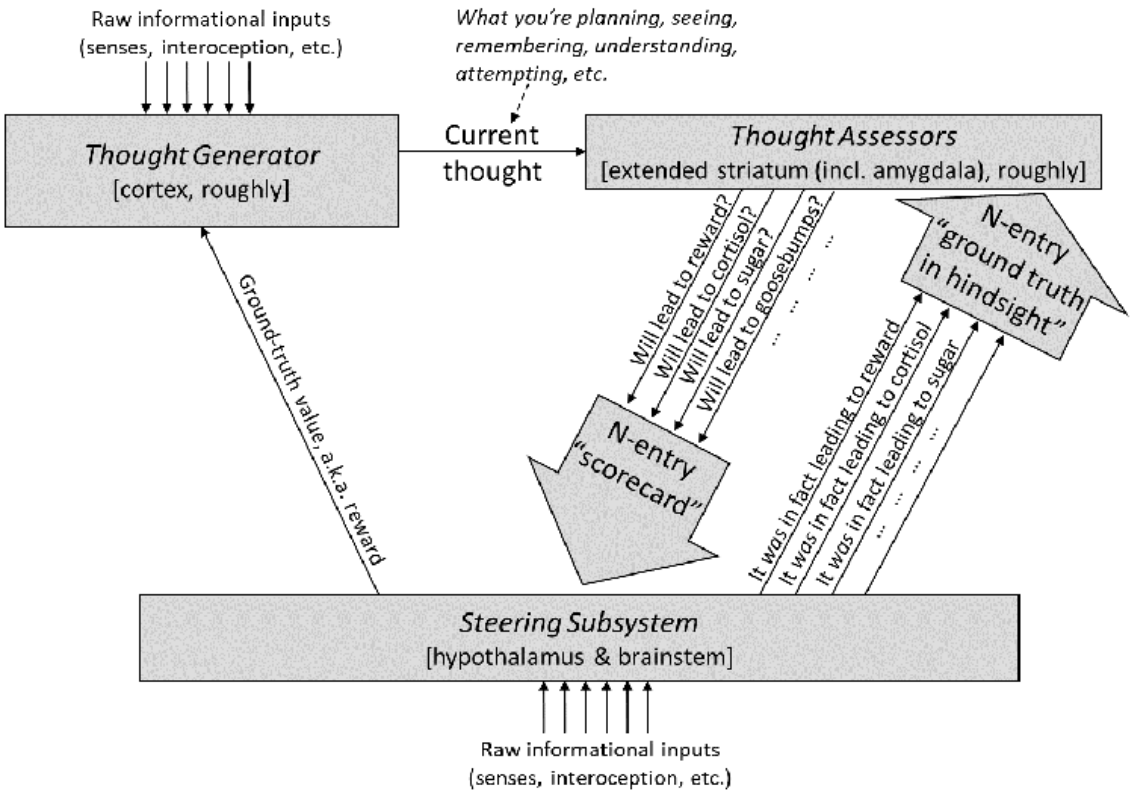
Third time seeing the no-longer-stranger:

- Steering Subsystem sensory heuristics say: “An adult human is present.”
- Thought Assessors say: “I expect to feel relaxed and playful and not-scared.”
- Steering Subsystem “Stranger Danger circuit” says: “Considering all of the above, we should be relaxed and playful and not-scared right now.”

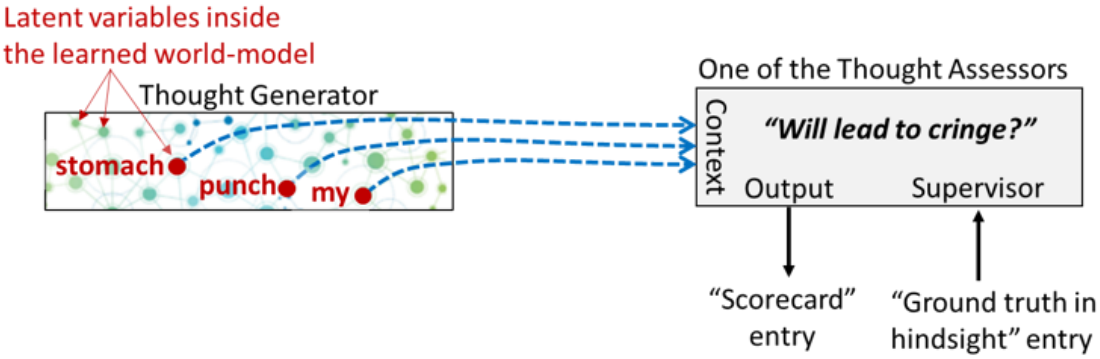
13.5 Another key ingredient (I think): “Little glimpses of empathy”

13.5.1 Introduction

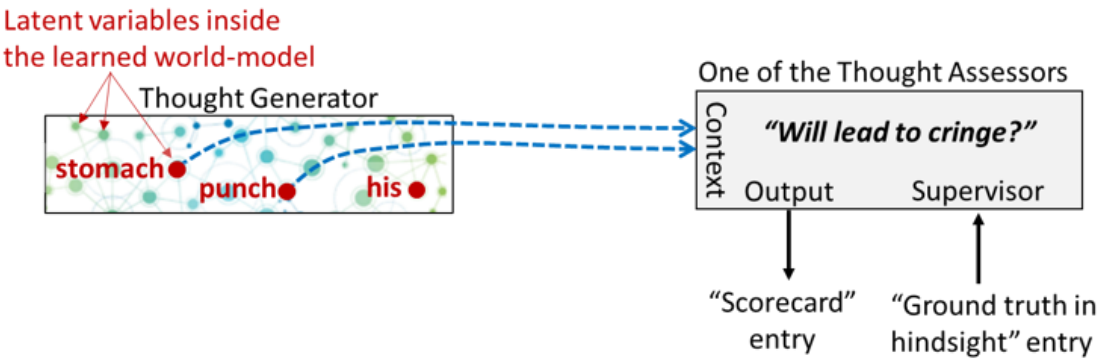
Yet again, here’s our diagram from Post #6:



Let's zoom in on one particular Thought Assessor in my brain, which happens to be dedicated to predicting a cringe reaction. This Thought Assessor has learned over the course of my lifetime that the predictive world-model activations corresponding to "my stomach is getting punched" constitute an appropriate time to cringe:



Now what happens when I watch someone else getting punched in the stomach?



If you look carefully on the left, you'll see that *"His stomach is getting punched"* is a different set of activations in my predictive world-model than *"My stomach is getting punched"*. But it's not *entirely* different! Presumably, the two sets would overlap to some degree.

And therefore, we should expect that, by default, "His stomach is getting punched" would send a weaker but nonzero "cringe" signal down to the Steering Subsystem.

I call this signal a "little glimpse of empathy". It tends to be a transient echo of what I (involuntarily) infer a different person to be feeling.

So what? Well, recall the symbol-grounding problem from Section 13.2.2 above. The existence of "little glimpses of empathy" is a massive breakthrough towards solving that problem for social instincts! After all, my Steering Subsystem now has a legible-to-it indication that a different person is feeling a certain feeling, and that signal can in turn trigger a response reaction *in me*.

(I'm glossing over various issues with "little glimpses of empathy", but I think those issues are solvable.[6])

For example, a (massively-oversimplified) envy reaction could look like "if I'm not happy, and I become aware (via a 'little glimpse of empathy') that someone else *is* happy, then issue a negative reward".

More generally, one could have a Steering Subsystem circuit whose inputs include:

1. my own current physiological state ("feelings"),
2. the contents of the "little glimpse of empathy",
3. …associated with some metadata about the person being empathetically simulated (maybe via a "perceived social status" Thought Assessor, for example?), and
4. heuristics drawn from my brainstem sensory processing systems, e.g. indicating whether I'm looking at a human right now.

The circuit could then produce outputs ("reactions"), which could (among other things) include rewards, other feelings, and/or ground truths for one or more Thought Assessors.

It seems to me that evolution would thus have quite a versatile toolbox for building social instincts, especially by chaining together more than one circuit of this type.

13.5.2 Distinction from the standard definition of "empathy"

I want to strongly distinguish "little glimpses of empathy" from the standard definition of "empathy".[7] (Maybe call the latter "a giant gulp of empathy"?)

For one thing, standard empathy is often effortful and voluntary, and may require at least a second or two of time, whereas a "little glimpse of empathy" is always fast and involuntary. An analogy for the latter would be how looking at a chair activates the "chair" concept in your brain, within a fraction of a second, whether you want it to or not.

For another thing, a "little glimpse of empathy", unlike standard "empathy", does not always lead to prosocial concern for its target. For example:

- In envy, if a little glimpse of empathy indicates that someone is happy, it makes me unhappy.
- In schadenfreude, if a little glimpse of empathy indicates that someone is unhappy, it makes me happy.
- When I'm angry, if a little glimpse of empathy indicates that the person I'm talking to is happy and calm, it sometimes makes me *even more* angry!

These examples are all antithetical to prosocial concern for the other person. Of course, in other situations, the "little glimpses of empathy" *do* spawn prosocial reactions. Basically, social instincts span the range from kind to cruel, and I suspect that pretty much all of them involve "little glimpses of empathy".

By the way: I already offered a model of "little glimpses of empathy" in the previous subsection. You might ask: What's my corresponding model of standard (giant gulp of) empathy?

Well, in the previous subsection, I distinguished "my own current physiological state (feelings)" from "the contents of the little glimpse of empathy". For standard empathy, I think this distinction breaks down—the latter bleeds into the former. Specifically, I would propose that when my Thought Assessors issue a sufficiently strong and long-lasting empathetic prediction, the Steering Subsystem starts "deferring" to them (in the Post #5 sense), and the result is that my own feelings wind up matching the feelings of the target-of-empathy. That's my model of standard empathy.

Then, if the target of my (standard) empathy is currently feeling an aversive feeling, I *also* wind up feeling an aversive feeling, and I don't like that, so I'm motivated to help him feel better (or, perhaps, motivated to shut him out, as can happen in compassion fatigue). Conversely, if the target of my (standard) empathy is currently feeling a pleasant feeling, I *also* wind up feeling a pleasant feeling, and I'm motivated to help him feel that feeling again.

Thus, standard empathy seems to be inevitably prosocial.

13.5.3 Why do I believe that "Little glimpses of empathy" are part of the story?

First, it seems introspectively right (to me, at least). If my friend is impressed by something I did, I feel proud, but I *especially* feel proud at the exact moment when I imagine my friend feeling that emotion. If my friend is disappointed in me, I feel guilty, but I *especially* feel guilty at the exact moment when I imagine my friend feeling that emotion. As another example, there's a saying: *"I can't wait to see the look on his face when…."* Presumably this saying reflects some real aspect of our social psychology, and if so, I claim that this observation dovetails well with my "little glimpses of empathy" story.

Second, way back in Post #5, Section 5.5.4, I noted that the medial prefrontal cortex (mPFC) (and the corresponding parts of the ventral striatum) plays a dual role as (1) a visceromotor center that can orchestrate autonomic reactions like pupil dilation and heart rate changes, and (2) a motivational / decision-making center. I claimed that the "Thought Assessors" picture elegantly explains why those roles go together as two sides of the same coin. I neglected to mention yet another role of mPFC, namely (3) a center of social instincts and morality. (Other Thought Assessor areas besides mPFC are in this category as well.) I think the "little glimpses of empathy" picture elegantly accounts for that as well: the "glimpses of empathy" correspond to signals getting sent from mPFC and the other Thought Assessor areas down to the Steering Subsystem, and thus all behavior that connects to social instincts necessarily involves Thought Assessors.

(That said, there are other possible social-instinct stories that *also* involve Thought Assessors but do *not* involve "little glimpses of empathy"—see for example Sections 13.3–13.4 above—so this piece of evidence is not very specific.)

Third, if the rest of my model (Posts #2–#7) is correct, then "little glimpse of empathy" signals would arise automatically, such that it would be straightforward to evolve a Steering Subsystem circuit that "listens" for them.

Fourth, if the rest of my model is correct, then, well, I can't think of any other way to build most social instincts! Process of elimination!

13.6 Future work (please!)

As noted in the introduction, the point of this post is to gesture towards what I expect a "theory of human social instincts" to look like, such that it would be compatible with all my other claims about brain algorithms in Posts #2–#7, particularly the strong constraint of "learning from scratch" as discussed in Section 13.2.2 above. My takeaway from the discussion in Sections 13.3–5 is a strong feeling of optimism that such a theory exists, even if I don't know all the details yet, and a corresponding optimism that this theory is actually how the human brain works, and will line up with corresponding circuits in the brainstem or (more likely) hypothalamus.

Of course, I want very much to move past the "general theorizing" stage, into more specific claims about how human social instincts actually work. For example, I'd love to move beyond speculation on how these instincts *might* solve the symbol-grounding problem, and learn how they *actually do* solve the symbol-grounding problem. I'm open to any ideas and pointers here, or better yet, for people to just figure this out on their own and tell me the answer.

For reasons discussed in the previous post, nailing down human social instincts is at the top of my wishlist for how neuroscientists can help with AGI safety.

Remember how I talked about Differential Technological Development (DTD) in Post #1 Section 1.7? Well, this is the DTD "ask" that I feel strongest about—at least, among those things that neuroscientists can do without explicitly working on AGI safety (see upcoming Post #15 for my more comprehensive wish-list). I *really* want us to reverse-engineer human social instincts in the hypothalamus & brainstem *long before* we reverse-engineer human world-modeling in the neocortex.

And things are not looking good for that project! The hypothalamus is small and deep and hence hard-to-study! Human social instincts might be different from rat social instincts! Orders of magnitude more research effort is going towards understanding neocortex world-modeling than understanding hypothalamus & brainstem social instinct circuitry! In fact, I've noticed (to my chagrin) that algorithmically-minded, AI-adjacent neuroscientists are *especially* likely to spend their talents on the Learning Subsystem (neocortex, hippocampus, cerebellum, etc.) rather than the hypothalamus & brainstem. But still, I don't think my DTD "ask" is hopeless, and I encourage anyone to try, and if you (or your lab) are in a good position to make progress but would need funding, email me and I'll keep you in the loop about possible upcoming opportunities.

1. ^

   See for example "The Evolutionary Psychology of Envy" by Hill & Buss, book chapter in *Envy: Theory & Research*, 2008.

2. ^

   Envy is on Donald E. Brown's "list of human universals", as reproduced in an appendix to *The Blank Slate* (Steven Pinker, 2002).

3. ^

   "…if you look at the human literature nobody talks about the hypothalamus and behaviour. The hypothalamus is very small and can't be readily seen by human brain imaging technologies like functional magnetic resonance imaging (fMRI). Also, much of the anatomical work in the instinctive fear system, for example, has been overlooked because it was carried out by Brazilian neuroscientists who were not particularly bothered to publish in high profile journals. Fortunately, there has recently been a renewed interest in these behaviors and these studies are being newly appreciated." (Cornelius Gross, 2018)

4. ^

   A few random example papers on the role of the Steering Subsystem (especially hypothalamus) in social behavior: "Independent hypothalamic circuits for social and predator fear" (Silva *et al.*, 2013), "Representation of distinct reward variables for self and other in primate lateral hypothalamus" (Noritake *et al.*, 2020), and "Social Stimuli Induce Activation of Oxytocin Neurons Within the Paraventricular Nucleus of the Hypothalamus to Promote Social Behavior in Male Mice" (Resendez *et al.*, 2020).

5. ^

   I suspect a more accurate diagram would feature arousal (in the psychology-jargon sense, not the sexual sense—i.e., heart rate elevation etc.) as a mediating variable. Specifically: (1) if brainstem sensory processing indicates that an adult human is present and nearby and picking me up etc., that leads to heightened arousal (by default, unless the Thought Assessors strongly indicate otherwise), and (2) when I'm in a state of heightened arousal, my brainstem treats it as bad and dangerous (by default, unless the Thought Assessors strongly indicate otherwise).

6. ^

   For example, the Steering Subsystem needs a method to distinguish a “little glimpse of empathy” from other transient feelings, e.g. the transient feeling that occurs when I think through the consequences of a possible course of action that I might take. Maybe there are some imperfect heuristics that could do that, but my preferred theory is that there’s a special Thought Assessor trained to fire when attending to another human (based on ground-truth sensory heuristics as discussed in Section 13.4). As another example, we need the “Ground truth in hindsight” signals to *not* gradually train away the Thought Assessor’s sensitivity to “his stomach is getting punched”. But it seems to me that, if the Steering Subsystem can figure out when a signal is a “little glimpse of empathy”, then it can choose not to send error signals to the Thought Assessors in those cases.

7. ^

   Warning: I’m not entirely sure that there really is a “standard” definition of empathy; it’s also possible that the term is used in lots of slightly-inconsistent ways.

[Intro to brain-like-AGI safety] 14. Controlled AGI

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

Part of the "Intro to brain-like-AGI safety" post series .

14.1 Post summary / Table of contents

Post #12 suggested two paths forward for solving "the alignment problem" for brain-like AGI, which I called "Social-instinct AGI" and "Controlled AGI". Then Post #13 went into more detail about (one aspect of) "Social-instinct AGI". And now, in this post, we're switching over to "Controlled AGI".

If you haven't read Post #12, don't worry, the "Controlled AGI" research path is nothing fancy—it's merely the idea of solving the alignment problem in the most obvious way possible:

The "Controlled AGI" research path:

- **Step 1 *(out-of-scope for this series)*:** We decide what we want our AGI's motivation to be. For example, that might be:
  - "Invent a better solar cell without causing catastrophe" (task-directed AGI),
  - "Be a helpful assistant to the human supervisor" (corrigible AGI assistants),
  - "Fulfill the human supervisor's deepest life goals" (ambitious value learning),
  - "Maximize coherent extrapolated volition",
  - or whatever else we choose.
- **Step 2 *(subject of this post)*:** We make an AGI with that motivation.

This post is about Step 2, whereas Step 1 is out-of-scope for this series. Honestly, I'd be *ecstatic* if we figured out how to reliably set the AGI's motivation to *any* of those things I mentioned under Step 1.

Unfortunately, I don't know any good plan for Step 2, and (I claim) nobody else does either. But I do have some vague thoughts and ideas, and I will share them here, in the spirit of brainstorming. This post is not meant to be a comprehensive overview of the whole problem, just what I see as the most urgent missing ingredients.

Out of all the posts in the series, this post is the hands-down winner for "most lightly-held opinions". For almost anything I say in this post, I can easily imagine someone changing my mind within an hour of conversation. Let that 'someone' be you—the comment section is below!

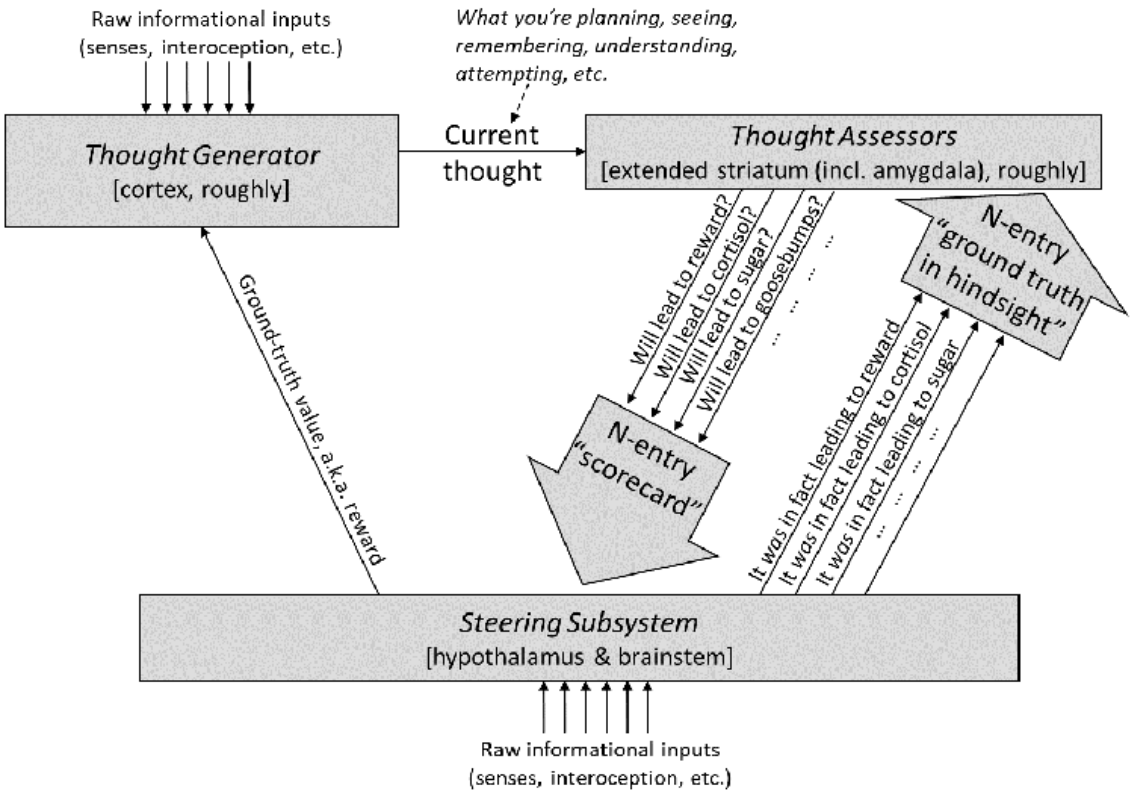
Table of contents:

- Section 14.2 discusses what we might use as "Thought Assessors" in an AGI. If you're just tuning in, Thought Assessors were defined in Posts #5–#6 and have been discussed throughout the series. If you have a Reinforcement Learning background, think of Thought Assessors as the components of a multi-dimensional value function. If you have a "being a human" background, think of Thought Assessors as learned functions that trigger visceral reactions (aversion, cortisol-release, etc.) based on the thought that you're consciously thinking right now. In the case of brain-like AGIs, we get to pick whatever Thought Assessors we want, and I propose three categories for consideration: Thought Assessors oriented towards safety (e.g. "this thought / plan

involves me being honest"), Thought Assessors oriented towards accomplishing a task (e.g. "this thought / plan will lead to better solar cell designs"), and Thought Assessors oriented purely towards interpretability (e.g. "this thought / plan has something to do with dogs").
- Section 14.3 discusses how we might generate supervisory signals to train those Thought Assessors. Part of this topic is what I call the "first-person problem", namely the open question of whether it's possible to take third-person labeled data (e.g. a YouTube video where Alice deceives Bob), and transmute it into a first-person preference (an AGI's desire to not, itself, be deceptive).
- Section 14.4 discusses the problem that the AGI will encounter "edge cases" in its preferences—plans or places where its preferences become ill-defined or self-contradictory. I'm cautiously optimistic that we can build a system that monitors the AGI's thoughts and detects when it encounters an edge case. However, I don't have any good idea about what to do when that happens. I'll discuss a few possible solutions, including "conservatism", and a couple different strategies for what Stuart Armstrong calls Concept Extrapolation.
- Section 14.5 discusses the open question of whether we can rigorously prove anything about an AGI's motivations. Doing so would seem to require diving into the AGI's predictive world-model (which would probably be a multi-terabyte, learned-from-scratch, unlabeled data structure), and proving things about what the components of the world-model "mean". I'm rather pessimistic about our prospects here, but I'll mention possible paths forward, including John Wentworth's "Natural Abstraction Hypothesis" research program (most recent update here).
- Section 14.6 concludes with my overall thoughts about our prospects for "Controlled AGIs". I'm currently a bit stumped and pessimistic about our prospects for coming up with a good plan, but hope I'm wrong and intend to keep thinking about it. I also note that a mediocre, unprincipled approach to "Controlled AGIs" would not *necessarily* cause a world-ending catastrophe—I think it's hard to say.

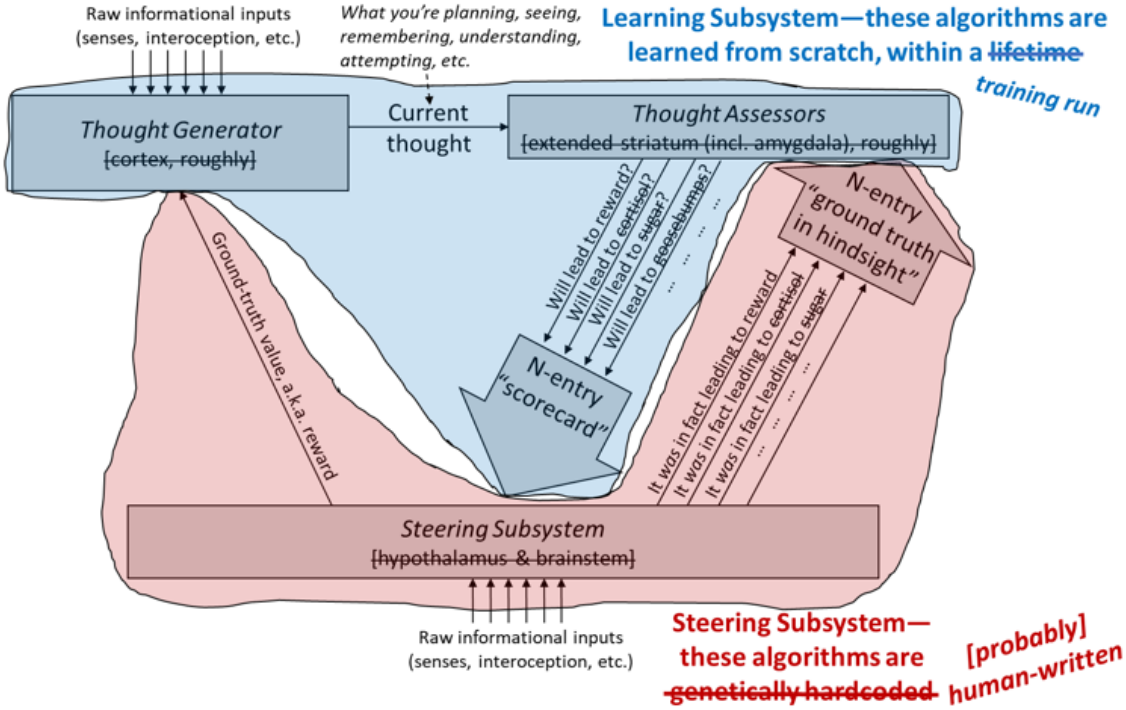
14.2 Three categories of AGI Thought Assessors

As background, here's our usual diagram of motivation in the human brain, from Post #6:



See Post #6.

And here's the modification for AGI, from Post #8:



On the center-right side of the diagram, I crossed out the words “cortisol”, “sugar”, “goosebumps”, etc. These correspond to the set of human innate visceral reactions which can be involuntarily triggered by thoughts (see Post #5). (Or in machine learning terms, these are more-or-less the components of a multidimensional value function, similar to what you find in multi-objective / multi-criteria reinforcement learning.)

Clearly, things like cortisol, sugar, and goosebumps are the wrong Thought Assessors for our future AGIs. But what are the right ones? Well, we’re the programmers! We get to decide!

I have in mind three categories to pick from. I’ll talk about how they might be trained (i.e., supervised) in Section 14.3 below.

14.2.1 Safety & corrigibility Thought Assessors

Example thought assessors in this category:

1. This thought / plan involves me being helpful.
2. This thought / plan does not involve manipulating my own learning process, code, or motivation systems.
3. This thought / plan does not involve deceiving or manipulating anyone.
4. This thought / plan does not involve anyone getting hurt.
5. This thought / plan involves following human norms, or more generally, doing things that an ethical human would plausibly do.
6. This thought / plan is “low impact” (according to human common sense).
7. …

Arguably (cf. this Paul Christiano post), #1 is enough, and subsumes the rest. But I dunno, I figure it would be nice to have information broken down on all these counts, allowing us to change the relative weights in real time (Post #9, Section 9.7), and perhaps giving an additional measure of safety.

Items #2–#3 are there because those are especially probable and dangerous types of thoughts—see discussion of Instrumental Convergence in Post #10, Section 10.3.2.

Item #5 is a bit of a catch-all for the AGI finding weird out-of-the-box solutions to problems, i.e. it’s my feeble attempt to mitigate the so-called “Nearest Unblocked Strategy problem”. Why might it mitigate the problem? Because pattern-matching to “things that an ethical human would plausibly do” is a *bit* more like a whitelist than a blacklist. I still don’t think that would work on its own, don't get me wrong, but *maybe* it would work in conjunction with the various other ideas in this post.

Before you jump into loophole-finding mode (*“lol an ethical human would plausibly turn the world into paperclips if they’re under the influence of alien mind-control rays”*), remember (1) these are meant to be implemented via pattern-matching to previously-seen examples (Section 14.3 below), not literal-genie-style following the exact words of the text; (2) we would hopefully also have some kind of out-of-distribution detection system (Section 14.4 below) to prevent the AGI from finding and exploiting weird edge-cases in that pattern-matching process. That said, as we’ll see, I don’t quite know how to do either of those two things, and even if we figure it out, I don’t have an airtight argument that it would be sufficient to get the intended safe behavior.

14.2.2 Task-related Thought Assessors

Example thought assessors in this category:

- This thought / plan will lead to a reduction in global warming

- This thought / plan will lead to a better solar panel design
- This thought / plan will lead to my supervisor becoming fabulously rich
- …

This kind of thing is why we built the AGI—what we actually want it to do. (Assuming task-directed AGI for simplicity.)

Basing a motivation system on these kinds of assessments *by themselves* would be obviously catastrophic. But maybe if we use these as motivations, *in conjunction with* the previous category, it will be OK. For example, imagine the AGI can only think thoughts that pattern-match to "I am being helpful" AND pattern-match to "there will be less global warming".

That said, I'm not sure we want this category at all. Maybe the "I am being helpful" Thought Assessor by itself is sufficient. After all, if the human supervisor is trying to reduce global warming, then a helpful AGI would produce a plan to reduce global warming. That's kinda the approach here, I think.

14.2.3 "Ersatz interpretability" Thought Assessors

(See Post #9, Section 9.6 for what I mean by "Ersatz interpretability".)

As discussed in Posts #4–#5, each thought assessor is a model trained by supervised learning. Certainly, the more Thought Assessors we put into the AGI, the more computationally expensive it will be. But I don't know *how much* more. Maybe we can put in 10^7 of them, and it only adds 1% to the total compute required by the AGI. I don't know. So I'll hope for the best and take the More Dakka approach: let's put in 30,000 Thought Assessors, one for every word in the dictionary:

- This thought / plan has something to do with AARDVARK
- This thought / plan has something to do with ABACUS
- This thought / plan has something to do with ABANDON
- … … …
- This thought / plan has something to do with ZOOPLANKTON

I expect that ML-savvy readers will be able to immediately suggest much-improved versions of this scheme—including versions with *even more* dakka—that involve things like contextual word embeddings and language models and so on. As one example, if we buy out and open-source Cyc (more on which below), we could use its hundreds of thousands of human-labeled concepts.

14.2.4 Combining Thought Assessors into a reward function

For an AGI to judge a thought / plan as being good, we'd like *all* the safety & corrigibility Thought Assessors from Section 14.2.1 to have as high a value as possible, *and* we'd like the task-related Thought Assessor from Section 14.2.2 (if we're using one) to have as high a value as possible.

(The outputs of the interpretability Thought Assessors from Section 14.2.3 are not inputs to the AGI's reward function, or indeed used at all in the AGI, I presume. I was figuring that they'd be silently spit out to help the programmers do debugging, testing, monitoring, etc.)

So the question is: how do we combine this array of numbers into a single overall score that can guide what the AGI decides to do?

A probably-bad answer is “add them up”. We don’t want the AGI going with a plan that performs catastrophically badly on all but one of the safety-related Thought Assessors, but so astronomically well on the last one that it makes up for it.

Instead, I imagine we’ll want to apply some kind of nonlinear function with strongly diminishing returns, and/or maybe even acceptability thresholds, before adding up the Thought Assessors into an overall score.

I don’t have much knowledge or opinion about the details. But there is some related literature on “scalarization” of multi-dimensional value functions—see here for some references.

14.3 Supervising the Thought Assessors, and the “first-person problem”

Recall from Posts #4–#6 that the Thought Assessors are trained by supervised learning. So we need a supervisory signal—what I labeled “ground truth in hindsight” in the diagram at the top.

I’ve talked about how the brain generates ground truth in numerous places, e.g. Post #3 Section 3.2.1, Posts #7 & #13. How do we generate it for the AGI?

Well, one obvious possibility is to have the AGI watch YouTube, with lots of labels throughout the video for when we think the various Thought Assessors ought to be active. Then when we’re ready to send the AGI off into the world to solve problems, we turn off the labeled YouTube videos, and simultaneously freeze the Thought Assessors (= set the error signals to zero) in their current state. Well, I’m not sure if that would work; maybe the AGI has to go back and watch more labeled YouTube videos from time to time, to help the Thought Assessors keep up as the AGI’s world-model grows and changes.

One potential shortcoming of this approach is related to first-person versus third-person concepts. We want the AGI to have strong preferences about aspects of first-person plans—hopefully, the AGI will see “I will lie and deceive” as bad, and “I will be helpful” as good. But we can’t straightforwardly get that kind of preference from the AGI watching labeled YouTube videos. The AGI will see YouTube character Alice deceiving YouTube character Bob, but that’s different from the AGI itself being deceptive. And it’s a very important difference! Consider:

- If you tell me “my AGI dislikes being deceptive”, I’ll say “good for you!”.
- If you tell me “my AGI dislikes it when people are deceptive”, I’ll say “for god's sake you better shut that thing off before it escapes human control and kills everyone”!!!

It sure would be great if there were a way to *transform* third-person data (e.g. a labeled YouTube video of Alice deceiving Bob) into an AGI’s first-person preferences (“I don’t want to be deceptive”). I call this **the first-person problem**.

How do we solve the first-person problem? I’m not entirely sure. Maybe we can apply interpretability tools to the AGI’s world-model, and figure out how it represents itself, and then correspondingly manipulate its thoughts, or something? It’s also possible that further investigation into human social instincts (previous post) will shed some light, as human social instincts do seem to transform the third-person “everyone in my friend group is wearing green lipstick” into the first-person “I want to be wearing green lipstick”.

If the first-person problem is not solvable, we need to instead use the scary method of allowing the AGI to take actions, and putting labels on those actions. Why is that scary? First, because those actions might be dangerous. Second, because it doesn't give us any good way to distinguish (for example) "the AGI said something dishonest" from "the AGI *got caught* saying something dishonest". Conservatism and/or concept extrapolation (Section 14.4 below) could help with that "getting caught" problem—maybe we could manage to get our AGI *both* motivated to be honest *and* motivated to not get caught, and that could be good enough—but it still seems fraught for various reasons.

14.3.1 Side note: do we *want* first-person preferences?

I suspect that "the first-person problem" is intuitive for most readers. But I bet a subset of readers feel tempted to say that the first-person problem is not in fact a problem at all. After all, in the realm of human affairs, there's a good argument that we could use a lot *fewer* first-person preferences!

The opposite of first-person preferences would be "impersonal consequentialist preferences", wherein there's a future situation that we want to bring about (e.g. "awesome post-AGI utopia"), and we make decisions to try to bring that about, *without* particular concern over what I-in-particular am doing. Indeed, too much first-person thinking leads to lots of things that I personally dislike in the world—e.g. jockeying for credit, blame avoidance, the act / omission distinction, social signaling, and so on.

Nevertheless, I still think giving AGIs first-person preferences is the right move for safety. Until we can establish super-reliable 12th-generation AGIs, I'd like them to treat "a bad thing happened (which had nothing to do with me)" as *much less bad* than "a bad thing happened (and it's my fault)". Humans have this notion, after all, and it seems at least *relatively* robust—for example, if I build a bank-robbing robot, and then it robs the bank, and then I protest "Hey I didn't do anything wrong; it was the robot!", I wouldn't be fooling anybody, much less myself. An AGI with such a preference scheme would presumably be cautious and conservative when deciding what to do, and would default to inaction when in doubt. That seems generally good, which brings us to our next topic:

14.4 Conservatism and concept-extrapolation

14.4.1 Why not just relentlessly optimize the right abstract concept?

Let's take a step back.

Suppose we build an AGI such that it has positive valence on the abstract concept "there will be lots of human flourishing", and consequently makes plans and take actions to make that concept happen.

I'm actually pretty optimistic that we'll be able to do that, from a technical perspective. Just as above, we can use labeled YouTube videos and so on to make a Thought Assessor for "this thought / plan will lead to human flourishing", and then base the reward function purely on that one Thought Assessor (cf. Post #7).

And then we set the AGI loose on an unsuspecting world, to go do whatever it thinks is best to do.

What could go wrong?

The problem is that the concept of "human flourishing" is an abstract concept in the AGI's world-model—really, it's just a fuzzy bundle of learned associations. It's hard to know what actions a desire for "human flourishing" will induce, especially as the world itself changes, and the AGI's understanding of the world changes even more. In other words, there is no future world that will *perfectly* pattern-match to the AGI's current notion of "human flourishing", and if an extremely powerful AGI optimized the world for the best possible pattern-match, we might wind up with something weird, even catastrophic. (Or maybe not! It's pretty hard to say, more on which in Section 14.6.)

As some random examples of what might go wrong: maybe the AGI would take over the world and prevent humans and human society from changing or evolving forevermore, because those changes would reduce the pattern-match quality. Or maybe the least-bad pattern-match would be the AGI wiping out actual humans in favor of an endless modded game of *The Sims*. Not that *The Sims* is a perfect pattern-match to "human flourishing"—it's probably pretty bad! But maybe it's less bad a pattern-match than anything the AGI could feasibly do with actual real-world humans. Or maybe as the AGI learns more and more, its world-model gradually drifts and changes, such that the frozen Thought Assessor winds up pointing at something totally random and crazy, and then the AGI wipes out humans to tile the galaxy with paperclips. I don't know!

So anyway, relentlessly optimizing a fixed, frozen abstract concept like "human flourishing" seems maybe problematic. Can we do better?

Well, it would be nice if we could also *continually refine* that concept, especially as the world itself, and the AGI's understanding of the world, evolves. This idea is what Stuart Armstrong calls Concept Extrapolation, if I understand correctly.

Concept extrapolation is easier said than done—there's no obvious ground truth for the question of "what is 'human flourishing', *really*?" For example, what would "human flourishing" mean in a future of transhuman brain-computer hybrid people and superintelligent evolved octopuses and god-only-knows-what-else?

Anyway, we can consider two steps to concept extrapolation. First (the easier part), we need to detect edge-cases in the AGI's preferences. Second (the harder part), we need to figure out what the AGI should do when it comes across such an edge-case. Let's talk about those in order.

14.4.2 The easier part of concept extrapolation: Detecting edge-cases in the AGI's preferences

I'm cautiously optimistic about the feasibility of making a simple monitoring algorithm that can watch an AGI's thoughts and detect that it's in an edge-case situation—i.e., an out-of-distribution situation where its learned preferences and concepts are breaking down.

(Understanding the *contents* of the edge-case seems much harder, as discussed shortly, but here I'm just talking about recognizing the *occurrence* of an edge-case.)

To pick a few examples of *possible* telltale signs that an AGI is at an edge-case:

- The learned probability distributions for Thought Assessors (see Post #5, Section 5.5.6.1) could have a wide variance, indicating uncertainty.
- The different Thought Assessors of Section 14.2 could diverge in new and unexpected ways.
- The AGI's reward prediction error could flip back and forth between positive and negative in a way that indicates "feeling torn" while attending to different aspects of the same possible plan.
- The AGI's generative world-model could settle into a state with very low prior probability, indicating confusion.

14.4.3 The harder part of concept extrapolation: What to do at an edge case

I don't know of any good answer. Here are some options.

14.4.3.1 Option A: Conservatism—When in doubt, just don't do it!

A straightforward approach would be that if the AGI's edge-case-detector fires, it forces the reward signal negative—so that whatever thought the AGI was thinking is taken to be a bad thought / plan. This would loosely correspond to a "conservative" AGI.

(Side note: I think there may be many knobs we can turn in order to make a brain-like AGI more or less "conservative", in different respects. The above is just one example. But they all seem to have the same issues.)

A failure mode of a conservative AGI is that the AGI just sits there, not doing anything, paralyzed by indecision, because every possible plan seems too uncertain or risky.

An "AGI paralyzed by indecision" is a failure mode, but it's not a *dangerous* failure mode. Well, not unless we were foolish enough to put this AGI in charge of a burning airplane plummeting towards the ground. But that's fine—in general, I think it's OK to have first-generation AGIs that can sometimes get paralyzed by indecision, and which are thus not suited to solving crises where every second counts. Such an AGI could still do important work like inventing new technology, and in particular designing better and safer second-generation AGIs.

However, if the AGI is *always* paralyzed by indecision—such that it can't get anything done—*now* we have a big problem. Presumably, in such a situation, future AGI programmers would just dial the "conservatism" knob down lower and lower, until the AGI started doing useful things. And at that point, it's unclear if the remaining conservatism would be sufficient to buy us safety.

I think it would be *much better* to have a way for the AGI to iteratively gain information to reduce uncertainty, while remaining highly conservative in the face of whatever uncertainty still remains. So how can we do that?

14.4.3.2 Option B: Dumb algorithm to seek clarification in edge-cases

Here's a slightly-silly illustrative example of what I have in mind. As above, we could have a simple monitoring algorithm that watches the AGI's thoughts, and detects when it's in an edge-case situation. As soon as it is, the monitoring algorithm shuts down the AGI entirely, and prints out the AGI's current neural net activations (and corresponding Thought Assessor outputs). The programmers use interpretability tools to figure out what the AGI is thinking about, and manually assign a value / reward, overriding the AGI's previous uncertainty with a highly-confident ground-truth.

That particular story seems unrealistic, mainly because we probably won't have sufficiently reliable and detailed interpretability tools. (Prove me wrong, interpretability researchers!) But maybe there's a better approach than just printing out billions of neural activations and corresponding Thought Assessors?

The tricky part is that AGI-human communication is fundamentally a hard problem. It's unclear to me whether it will be possible to solve that problem via a dumb algorithm. The situation here is very different from, say, an image classifier, where we can find an edge-case picture and just show it to the human. The AGI's thoughts may be much more inscrutable than that.

By analogy, human-human communication is possible, but not by any dumb algorithm. We do it by leveraging the full power of our intellect—modeling what our conversation partner is thinking, strategically choosing words that will best convey a desired message, and learning through experience to communicate more and more effectively. So what if we try that approach?

14.4.3.3 Option C: The AGI *wants* to seek clarification in edge-cases

If I'm trying to help someone, I don't need any special monitoring algorithm to prod me to seek clarification at edge-cases. Seeking clarification at edge-cases is just *what I want to do*, as a self-aware properly-motivated agent.

So what if we make our AGIs like that?

At first glance, this approach would seem to solve all the problems mentioned above. Not only that, but the AGI can use its full powers to make everything work better. In particular, it can learn its own increasingly-sophisticated metacognitive heuristics to flag edge-cases, and it can learn and apply the human's meta-preferences about how and when the AGI should ask for clarification.

But there's a catch. I was *hoping* for a conservatism / concept extrapolation system that would help protect us from misdirected motivations. If we implement conservatism / concept extrapolation via the motivation system itself, then we lose that protection.

More specifically: if we go up a level, the AGI *still* has a motivation ("seek clarification in edge-cases"), and that motivation is *still* an abstract concept that we have to extrapolate into out-of-distribution edge cases ("What if my supervisor is drunk, or dead, or confused? What if I ask a leading question?"). And for *that* concept extrapolation problem, we're plowing ahead without a safety net.

Is that a problem? Bit of a long story:

Side-debate: Will "helpfulness"-type preferences "extrapolate" safely just by recursively applying to themselves?

In fact, a longstanding debate in AGI safety is whether these kinds of helpful / corrigible AGI preferences (e.g. an AGI's desire to understand and follow a human's preferences and meta-preferences) will "extrapolate" in a desirable way without any "safety net"—i.e., without any *independent* ground-truth mechanism pushing the AGI's preferences in the right direction.

In the optimistic camp is Paul Christiano, who argued in "Corrigibility" (2017) that there would be "a broad basin of attraction towards acceptable outcomes", based on, for example, the idea that an AGI's preference to be helpful will result in the AGI having a self-reflective desire to continually edit its own preferences in a direction humans would like. But I don't really buy that argument for reasons in my 2020 post—basically, I think there are bound to be sensitive areas like "what does it mean for people to want something" and "what are human communication norms" and "inclination to self-monitor", and if the AGI's preferences drift along any of those axes (or all of them simultaneously), I'm not convinced that those preferences would self-correct.

Meanwhile, in the strongly-pessimistic camp is Eliezer Yudkowsky, I think mainly because of an argument (e.g. this post, final section) that we should expect powerful AGIs to have consequentialist preferences, and that consequentialist preferences seem incompatible with corrigibility. But I don't really buy that argument either, for reasons in my 2021 "Consequentialism & Corrigibility" post—basically, I think there are possible preferences that are reflectively-stable, and that *include* consequentialist preferences (and thus are compatible with powerful capabilities), but are not *purely* consequentialist (and thus are compatible with corrigibility). A "preference to be helpful" seems like it could plausibly develop into that kind of hybrid preference scheme.

Anyway, I'm uncertain but leaning pessimistic. For more on the topic, see also Wei Dai's recent post, and the comment sections of all of the posts linked above.

14.4.3.4 Option D: Something else?

I dunno.

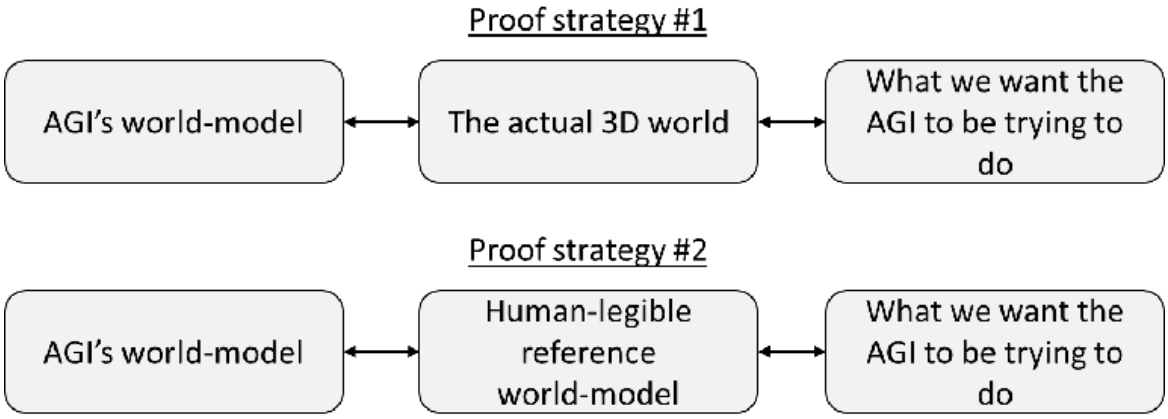
14.5 Getting a handle on the world-model itself

The elephant in the room is the giant multi-terabyte unlabeled generative world-model that lives inside the Thought Generator. The Thought Assessors provide a window into this world-model, but I'm concerned that it may be a rather small, foggy, and distorted window. Can we do better?

Ideally, we'd like to prove things about the AGI's motivation. We'd like to say "Given the state of the AGI's world-model and Thought Assessors, the AGI is definitely motivated to do X" (where X=be helpful, be honest, not hurt people, etc.) Wouldn't that be great?

But we immediately slam into a brick wall: How do we prove *anything whatsoever* about the "meaning" of things in the world-model, and thus about the AGI's motivation? The world is complicated, and therefore the world-model is complicated. The things we care about are fuzzy abstractions like "honesty" and "helpfulness"—see the Pointers Problem. The world-model keeps changing as the AGI learns more, and as it makes plans that would entail taking the world wildly out-of-distribution (e.g. planning the deployment of a new technology). How can we possibly prove anything here?

I still think the most likely answer is "We can't". But here are two possible paths anyway. For some related discussion, see Eliciting Latent Knowledge.



Proof strategy #1 starts with the idea that we live in a three-dimensional world containing objects and so on. We try to come up with an unambiguous definition of what those objects are, and from there we can have an unambiguous language for specifying what we want to happen in the world. We also somehow translate (or constrain) the AGI's understanding of the world into that language, and now we can prove theorems about what the AGI is trying to do.

This is my tentative understanding of what John Wentworth is trying to do via his Natural Abstraction Hypothesis research program (most recent update here), and I've heard ideas in this vicinity from a couple other people as well. (Update: John disagrees with this characterization, see his comment.)

I'm skeptical because a 3D world of localized objects seems to be an unpromising starting point for stating and proving useful theorems about the AGI's motivations. After all, a lot of things that we humans care about, and that the AGI needs to care about, seem difficult to describe in terms of a 3D world of localized objects—consider the notion of "honesty", or "solar cell efficiency", or even "daytime".

Proof strategy #2 would start with a human-legible "reference world-model" (e.g. Cyc). This reference world-model wouldn't be constrained to be built out of localized objects in a 3D world, so unlike the above, it could and probably would contain things like "honesty" and "solar cell efficiency" and "daytime".

Then we try to directly match up things in the "reference world-model" with things in the AGI's world-model.

Will they match up? No, of course not. Probably the best we can hope for is a fuzzy, many-to-many match, with various holes on both sides.

It's hard for me to see a path to rigorously proving anything about the AGI's motivations using this approach. Nevertheless, I continue to be amazed that unsupervised machine translation is possible at all, and I take that as an indirect hint that *if* pieces of two world-models match up with each other in their internal structure, *then* those pieces are probably describing the same real-world thing. So maybe I have the faintest glimmer of hope.

I'm unaware of work in this direction, possibly because it's stupid and doomed, and also possibly because I don't think we currently have any really great open-source *human-legible* world-models to run experiments on. The latter is a problem that I think we should rectify ASAP, perhaps by cutting a giant check to open-source Cyc, or else developing a similarly rich, accurate, and (most importantly) human-legible open-source world-model by some other means.

14.6 Conclusion: mild pessimism about finding a good solution, uncertainty about the consequences of a lousy solution

I think we have our work cut out figuring out how to solve the alignment problem via the "Controlled AGIs" route (as defined in Post #12). There are a bunch of open problems, and I'm currently pretty stumped. We should absolutely keep looking for good solutions, but right now I'm also open-minded to the possibility that we won't find any. That's why I continue to put a lot of my mental energy into the "social-instinct AGIs" path (Posts #12–#13), which seems somewhat less doomed to me, despite its various problems.

I note, however, that my pessimism is not universally shared—for example, as mentioned, Stuart Armstrong at AlignedAI appears optimistic about solving the open problem in Section 14.4, and John Wentworth appears optimistic about solving the open problem in Section 14.5. Let's hope they're right, wish them luck, and try to help!

To be clear, the thing I'm feeling pessimistic about is finding a *good* solution to "Controlled AGI", i.e., a solution that we can feel extremely confident in *a priori*. A different question is: Suppose we try to make "Controlled AGI" via a *lousy* solution, like the Section 14.4.1 example where we imbue a super-powerful AGI with an all-consuming desire for the abstract concept of "human flourishing", and the AGI then extrapolates that abstract concept arbitrarily far out of distribution in a totally-uncontrolled, totally-unprincipled way. Just how bad a future would such an AGI bring about? I'm very uncertain. Would such an AGI engage in mass torture? Umm, I guess I'm cautiously optimistic that it wouldn't, absent a sign error from cosmic rays or whatever. Would it wipe out humanity? I think it's possible!—see discussion in Section 14.4.1. But it might not! Hey, maybe it would even bring about a pretty awesome future! I just really don't know, and I'm not even sure how to reduce my uncertainty.

In the next post, I will wrap up the series with my wish-list of open problems, and advice on how to get into the field and help solve them!

[Intro to brain-like-AGI safety] 15. Conclusion: Open problems, how to help, AMA

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

15.1 Post summary / Table of contents

This is the final post of the *"Intro to brain-like-AGI safety"* post series! Thanks for reading this far!

- In Section 15.2, I'll list seven open problems that came up in the previous posts. I'm putting them all here in one place for the convenience of potential researchers and funders.
- In Section 15.3, I'll offer some brief remarks on practical aspects of doing AGI safety (a.k.a. AI alignment) research, including funding sources, connecting to the relevant research community, and where to learn more.
- In Section 15.4, I'll wrap up with 8 takeaway messages that I hope readers will have gotten out of this series.

Since this is the "Conclusion" post, feel free to use the comment section for more general discussion (or to "ask me anything"), even if it's not related to this particular post.

15.2 Open problems

This is not, by any stretch of the imagination, a complete list of open problems whose progress would help with brain-like-AGI safety, let alone with the more general topic of Safe & Beneficial AGI (see Post #1, Section 1.2). Rather, these are just some of the topics that came up in this series, with ratings proportional to how enthusiastic I am about them.

I'll split the various open problems into three categories: "Open problems that look like normal neuroscience", "Open problems that look like normal computer science", and "Open problems that require explicitly talking about AGIs". This division is for readers' convenience; you might, for example, have a boss, funding source, or tenure committee who thinks that AGI Safety is stupid, and in that case you might want to avoid the third category. (However, don't give up so soon—see discussion in Section 15.3.1 below.)

15.2.1 Open problems that look like normal neuroscience

15.2.1.1 The *"Is Steve full of crap when he talks about neuroscience?"* research program — ⭐⭐⭐⭐

If you didn't notice, Posts #2–#7 are full of grand theorizing and bold claims about how the human brain works. It would be nice to know if those claims are actually true!!

If those neuroscience posts are a bunch of baloney, then I think we should throw out not only those posts, but the whole rest of this series too.

In the text of those posts, you'll see various suggestions and pointers as to why I believe the various neuroscience claims that I made. But a careful, well-researched analysis has yet to be written, as far as I'm aware. (Or if it has, send me a link! Nothing would make me happier than learning that I'm reinventing the wheel by saying things that are already well-established and widely-accepted.)

I give this research program a priority score of **4 stars out of 5**. Why not 5? Two things:

- It loses half a star because I have utterly-unjustifiable overconfidence that my neuroscience claims are not, in fact, a bunch of baloney, and therefore this research program would look more like nailing down some of the finer details, and less like throwing this whole post series in the garbage.
- It loses another half star because I think there are some delicate corners of this research program where it gets uncomfortably close to the "unravel the gory details of the brain's learning-from-scratch algorithms" research program, a research program to which I assign *negative* 5 stars, because I'd like to make more progress on how and whether we can safely use a brain-like AGI, long before we figure out how to build one. (See Differential Technology Development discussion in Post #1, Section 1.7.)

15.2.1.2 The *"Reverse-engineer human social instincts"* research program — ⭐⭐⭐⭐⭐

Assuming that Posts #2–#7 are not, in fact, a bunch of baloney, the implication is that there are circuits for various "innate reactions" that underlie human social instincts, they are located somewhere in the "Steering Subsystem" part of the brain (roughly the hypothalamus and brainstem), and they are relatively simple input-output functions. The goal: figure out exactly what those input-output functions are, and how they lead (after within-lifetime learning) to our social and moral thoughts and behaviors.

See Post #12 for why I think this research program is very good for AGI safety, and Post #13 for more discussion of roughly what kinds of circuits and explanations we should be looking for.

Here's a (somewhat caricatured) more ML-oriented perspective on this same research program: It's widely agreed that the human brain within-lifetime learning algorithm involves reinforcement learning (RL)—for example, after you touch the hot stove once, you don't do it again. As with any RL algorithm, we can ask two questions:

1. How does the brain's RL algorithm work?
2. What exactly is the reward function?

These questions are (more-or-less) independent. For example, to study question A experimentally, you don't need a full answer to question B; all you need is at least one way to create a positive reward, and at least one way to create a negative reward, to use in your experiments. That's easy: Rats like eating cheese, and rats dislike getting electrocuted. Done!

My impression is that neuroscientists have produced many thousands of papers on question A, and practically none directly addressing question B. But I think question B is much *more* important for AGI safety. And the social-instincts-related parts of the reward function, which are upstream of morality-related intuitions, are most important of all.

I give this research program a priority score of **5 stars out of 5**, for reasons discussed in Posts #12–#13.

15.2.2 Open problems that look like normal computer science

15.2.2.1 The *"Make the biggest and best open-source human-legible world-model / web-of-knowledge that we can"* research program — ⭐⭐⭐

I first talked about this in a post "Let's buy out Cyc, for use in AGI interpretability systems?" (Despite the post title, I'm not overly tied to Cyc in particular; if today's machine learning magic can get the same job done better and cheaper, that's great.)

I expect that future AGIs will build and continually expand their own world-models, and those world-models will eventually grow to terabytes of information and beyond, and will include brilliant innovative concepts that humans have never thought of, and can't understand without years of study (or at all). Basically, we'll have our work cut out in making sense of an AGI's world-model. So what do we do? (No, "run away screaming" isn't an option.) It seems to me that if we have our own giant *human-legible* world-model, that would be a powerful tool in our arsenal as we attack the problem of understanding the AGI's world-model. The bigger and better the human-legible world-model, the more helpful it would be.

To be more specific, in previous posts I've mentioned **three reasons that having a huge, awesome, open-source human-legible world-model might be helpful**:

- *For non-learning-from-scratch initialization*—see Post #11, Section 11.3.1. By default, I expect that an AGI's world-model and Thought Assessors (roughly, RL value function) will be "learned from scratch" in the Post #2 sense. That means that an "infant AGI" will be thrashing around in the best case, and doing dangerous planning against our interests in the worst case, as we try to sculpt its preferences in a human-friendly direction. It would be awfully nice if we could *not* initialize from scratch, so as to avoid that problem. It's far from clear to me that a non-learning-from-scratch approach will be possible at all, but if it is, having a huge awesome human-legible world-model at our disposal would presumably help.
- *As a list of concept labels for "ersatz interpretability"*—see Post #14, Section 14.2.3. Cyc, for example, has hundreds of thousands of concepts, which are considerably more specific than English-language words—for example, a single word with 10 definitions would get split into 10 Cyc concepts with 10 different names. If we have a nice concept-list like that, and we have a bunch of labeled examples, then we can use supervised learning (or more simply, cross-correlation) to look for signs that particular patterns of AGI neural net activations are related to that AGI "thinking about" certain concepts.
- *As a "reference world-model" for "real" (or even rigorous) interpretability*—see Post #14, Section 14.5. This would involve digging deeper into both an AGI's world-model and the open-source human-legible "reference world-model", finding areas of deep structural similarity that overlap with the cross-correlations mentioned above, and inferring that these are really talking about the same aspects of the world. As discussed in that post, I give this a low probability of success (related: discussion of "ontology mismatches" here), but extremely high reward if it does succeed.

I give this research program a priority score of **3 stars out of 5**, because I don't have *super* high confidence that any of those three stories are both real and extremely impactful. I dunno, maybe there's a 50% chance that, even if we had a super-awesome open-source human-legible world-model, future AGI programmers wouldn't wind up using it, or else that it would only be marginally better than a *mediocre* open-source human-legible world-model.

15.2.2.2 The *"Easy-to-use super-secure sandbox for AGIs"* research program — ⭐⭐⭐

Recall from above: By default, I expect that an AGI's world-model and Thought Assessors (roughly, RL value function) will be "learned from scratch" in the Post #2 sense. That means that an "infant AGI" will be thrashing around in the best case, and doing dangerous planning against our interests in the worst case, as we try to sculpt its preferences in a human-friendly direction.

Given that, it would be nice to have a super-secure sandbox environment in which the "infant AGI" can do whatever learning it needs to do without escaping onto the internet or otherwise causing chaos.

Some possible objections:

- *Possible Objection #1: A perfectly secure sandbox is not realistic.* That might be true, I dunno. But I'm not talking about security against a superintelligent AGI, but rather against an "infant AGI" whose motivations and understanding of the world are still in flux. In that context, I think a more-secure sandbox is meaningfully better than a less-secure sandbox, even if neither is perfect. By the time the AGI is powerful enough to escape any imperfect sandbox, we'll have already (hopefully!) installed in it the motivation not to do so.
- *Possible Objection #2: We can already make a reasonably (albeit imperfectly) secure sandbox.* Again, that might be true; I wouldn't know either way. But I'm especially interested in whether future AGI programmers *will actually use* the best secure sandbox that we can build, under deeply cynical assumptions about the motivation and security skills of those programmers. (Related: "alignment tax".) That means that the super-secure sandbox needs to be polished, to be decked out with every feature that anyone could possibly want, to be user-friendly, to carry negligible performance penalty, and to be compatible with every aspect of how programmers actually train and run massive machine learning jobs. I suspect that there's room for improvement on all these counts.

I give this research program a priority score of **3 stars out of 5**, mostly because I don't know that much about this topic, and therefore I don't feel comfortable being its outspoken champion.

15.2.3 Open problems that involve explicitly talking about AGIs

15.2.3.1 The *"Edge-cases / conservatism / concept extrapolation"* research program — ⭐⭐⭐⭐⭐

Humans can easily learn the meaning of abstract concepts like "being a rock star", just by observing the world, pattern-matching to previously-seen examples, etc. Moreover, having learned that concept, humans can *want* (assign positive valence to) that concept, mainly as a result of repeatedly getting reward signals while that concept was active in their mind (see Post #9, Section 9.3). This seems to suggest a general strategy for controlling brain-like AGIs: prod the AGIs to learn particular concepts like "being honest" and "being helpful" via labeled examples, and then ensure that those concepts get positive valence, and then we're done!

However, concepts are built out of a web of statistical associations, and as soon as we go to out-of-distribution edge-cases, those associations break down, and so does the concept. If there's a religious fundamentalist who believes in a false god, are you being "helpful" if you

deconvert them? The best answer is “I don’t know, it depends on exactly what you mean by ‘helpful’”. Such an action matches well to *some* of the connotations / associations of the “helpfulness” concept, but matches quite poorly to other connotations / associations.

So prodding the AGI to learn and like certain abstract concepts seems like the *start* of a good plan, but only if we have a principled approach to making the AGI refine those concepts, in a way we endorse, upon encountering edge-cases. And here, I don’t have any great ideas.

See Post #14, Section 14.4 for further discussion.

Side note: If you’re really motivated by this research program, one option might be applying for a job at AlignedAI. Their co-founder Stuart Armstrong originally suggested “concept extrapolation” as a research program (and coined the term), and I believe that this is their main research focus. Given Stuart Armstrong’s long history of rigorous thinking about AGI safety, I’m cautiously optimistic that AlignedAI will work towards solutions that will scale to the superintelligent AGIs of tomorrow, instead of just narrowly targeting the AI systems of today, as happens far too often.

I give this research program a priority score of **5 stars out of 5**. Solving this problem would get us at least much of the way towards knowing how to build “Controlled AGIs” (in the Post #14 sense).

15.2.3.2 The *“Rigorously prove anything whatsoever about the meaning of things in a learned-from-scratch world-model”* research program — ⭐⭐⭐⭐⭐

The brain-like AGI will presumably learn-from-scratch a giant multi-terabyte unlabeled generative world-model. The AGI’s goals and desires will all be defined in terms of the contents of that world-model (Post #9, Section 9.2). And ideally, we’d like to make confident claims, or better yet prove theorems, about the AGI’s goals and desires. Doing so would seem to require proving things about the “meaning” of the entries in this complicated, constantly-growing world-model. How do we do that? I don’t know.

See discussion in Post #14, Section 14.5.

There’s some work in this general vicinity at Alignment Research Center, which does excellent work and is hiring. (See the discourse on ELK.) But as far as I know, making progress here is a hard problem that needs new ideas, if it’s even possible.

I give this research program a priority score of **5 stars out of 5**. Maybe it’s intractable, but it sure as heck would be impactful. It would, after all, give us complete confidence that we understand what an AGI is trying to do.

15.2.3.3 The *“Solving the whole problem”* research program — ⭐⭐⭐⭐⭐

This is the sort of thing I was doing in Posts #12 and #14. We need to tie everything together into a plausible story, figure out what’s missing, and crystallize how to move forward. If you read those posts, you’ll see that there’s a lot of work yet to do—for example, we need a much better plan for training data / training environments, and I didn’t even *mention* important ingredients like sandbox test protocols. But many of the design considerations seem to be interconnected, such that I can’t easily split it out into multiple different research programs. So this is my catch-all category for all that stuff.

(See also: Research productivity tip: “Solve The Whole Problem Day”.)

I give this research program a priority score of **5 stars out of 5**, for obvious reasons.

15.3 How to get involved

(Warning: this section may become rapidly out-of-date. I'm writing in May 2022.)

15.3.1 Funding situation

If you care about AGI safety (a.k.a. "AI alignment"), and your goal is to help with AGI safety, it's *extremely* nice to get funding from a funding source that has the same goal.

Of course, it's also possible to get funding from more traditional sources, e.g. government science funding, and use it in an AGI-safety-promoting way. But then you have to strike a compromise between "things that would help AGI safety" and "things that would impress / satisfy the funding source". My advice and experience is that this kind of compromise is really bad. I spent some time exploring this kind of compromise strategy early on in my journey into AGI safety; I had been warned that it was bad, and I *still* dramatically underestimated just how bad it was. If it's any indication, I wound up hobby-blogging about AGI safety in little bits of free time squeezed between a full-time job and two young kids, and I think that was *dramatically* more useful than if I had devoted all day every day to my best available "compromise" project.

(You can replace "compromise in order to satisfy my funding source" with "compromise in order to satisfy my thesis committee", or "compromise in order to satisfy my boss", or "compromise in order to have an impressive CV for my future job search / tenure review", etc., as appropriate.)

Anyway, as luck would have it, there are numerous funding sources that are explicitly motivated by AGI safety. They're all philanthropic foundations, as far as I'm aware. (I guess worrying about future out-of-control AGIs is just a bit too exotic for government funding agencies?) Funding for technical AGI safety (the topic of this series) has been growing rapidly, and seems to be in the tens of millions of dollars a year right now, maybe, depending in large part on your own particular spicy hot take about what does or doesn't count as *real* technical AGI safety research.

Many but not all AGI-safety-concerned philanthropists (and researchers like myself) are connected to the Effective Altruism (EA) movement, a community / movement / project devoted to trying to work out how best to make the world a better place, and then go do it. Within EA is a "longtermism" wing, consisting of people acting out of concern for the long-term future, where "long term" might mean millions or billions or trillions of years. Longtermists tend to be *especially* motivated to prevent irreversible human-extinction-scale catastrophes like out-of-control AGIs, bio-engineered pandemics, etc. Thus, in EA circles, AGI safety is sometimes referred to as a "longtermist cause area", which is kinda disorienting given that we're talking about how to prevent a potential calamity that could well happen in my lifetime (see timelines discussion in Posts #2–#3). Oh well.

(This is just lighthearted humor, not making fun of anyone—in fact, I myself am acting partly out of concern for the long-term future.)

The connection between EA and AGI safety has become sufficiently strong that (1) some of the best conferences to go to as an AGI safety researcher are the EA Global / EAGx conferences, and (2) people started calling me an EA, and cold-emailing me to invite me to EA events, totally unprompted, for the sole reason that I had recently started blogging about AGI safety in my free time.

Anyway, the point is: AGI-safety-motivated funding exists—whether you're in academia, in a nonprofit, or just an independent researcher (like me!). How do you get it? By and large, you probably need to either:

1. Demonstrate that you personally understand the AGI safety problem well enough to have good judgment about what research would be helpful, or
2. Jump onto a concrete research program that AGI-safety-experts have *already* endorsed as being important and useful.

As for #2, one reason that Section 15.2 exists is that I'm trying to help this process along. I imagine that at least some of those seven research programs above could (with some work) be fleshed out into a nice, specific, funded Request For Proposals. Email me if you think you could help, or want me to keep you in the loop.

As for #1—Yeah, go for it!! AGI safety is a fascinating field (IMHO), and it's sufficiently "young" that you can get up to the research frontier much faster than would be possible in, for example, particle physics. See the next subsection for links to resources, training courses, etc. Or I guess you can learn the field by reading and writing lots of blog posts and comments in your free time, like I did.

By the way, it's true that the nonprofit sector *in general* has a reputation for shoestring budgets and underpaid, overworked employees. But philanthropy-funded AGI safety work is generally not like that. The funders want the best people, even if those people are well into their careers and saddled with mortgage payments, daycare costs, etc.—like yours truly! So there has been a strong movement towards salaries that are competitive with the for-profit sector, especially in the past couple years.

15.3.2 Jobs, organizations, training programs, community, etc.

15.3.2.1 …For AGI safety (a.k.a. AI alignment) in general

There are lots of links at the aptly-named AI Safety Support Lots-of-Links page, or you can find a more-curated list at "AI safety starter pack". To call out just a couple particularly relevant items:

- 80,000 hours is an organization devoted to helping people do good through their careers. They're *very* into AGI safety, and they offer free 1-on-1 career counseling, in which they'll tell you about relevant opportunities and connect you to relevant people. Also check out their AI safety guide, the AI-technical-safety-related episodes of their excellent podcast, and their AI-specific email list and job board. (You can also get free 1-on-1 career coaching through AI Safety Support, no application needed.)
- You might be reading this article on lesswrong.com, a blogging platform which has the (I think) unique feature of being simultaneously open to anyone and frequented by numerous AGI safety experts. I started blogging and commenting there when I was just starting out in my free time in 2019, and I recall finding everyone very kind and helpful, and I don't know how else I could have gotten into the field, given my geographical and time constraints. Other active online congregation points include the EleutherAI discord, Robert Miles's discord, and AI Safety Support Slack. As for *in-person* local meetups / reading groups / etc., check here or here, or better yet stop by your local / university EA group and ask them for pointers.

15.3.2.2 …More specifically related to this series

Q: Is there a community gathering place for discussing "brain-like AGI safety" (or closely-related "model-based RL AGI safety") in particular?

A: Not really. And I'm not entirely sure that there should be, since it overlaps so much with other lines of research within AGI safety.

(The closest thing to that is maybe the discord server associated with so-called "shard theory", email me for the link.)

Q: Is there a community gathering place for discussing the overlap between neuroscience / psychology, and AGI safety / AI alignment?

A: There's a "neuroscience & psychology" channel in the AI Safety Support Slack. You can also join the email list for PIBBSS, in case that happens again in the future.

If you want to see more different perspectives in the neuroscience / AGI safety overlap area, check out papers by Kaj Sotala; Seth Herd, David Jilk, Randall O'Reilly *et al.*; Gopal Sarma & Nick Hay; Patrick Butlin; Jan Kulveit; along with other articles by those same authors, and many others that I'm rudely forgetting.

(My own background, for what it's worth, is in physics, not neuroscience—in fact, I knew essentially no neuroscience as recently as 2019. I got interested in neuroscience to help answer my burning questions related to AGI safety, not the other way around.)

Q: Hey Steve, can I work with you?

A: While I'm not currently interested in hiring or supervising anyone, I am always very happy to collaborate and correspond. There's plenty of work to do! Email me if you want to chat!

15.4 Conclusion: 8 takeaway messages

Thanks for reading! I hope that, in this series, I have successfully conveyed the following messages:

- We know enough neuroscience to say concrete things about what "brain-like AGI" would look like (Posts #1–#9);
- In particular, while "brain-like AGI" would be different from any known algorithm, its safety-relevant aspects would have much in common with actor-critic model-based reinforcement learning with a multi-dimensional value function (Posts #6, #8, #9);
- "Understanding the brain well enough to make brain-like AGI" is a dramatically easier task than "understanding the brain" full stop—if the former is loosely analogous to knowing how to train a ConvNet, then the latter would be loosely analogous to knowing how to train a ConvNet, *and* achieving full mechanistic interpretability of the resulting trained model, *and* understanding every aspect of integrated circuit physics and engineering, etc. Indeed, making brain-like AGI should not be thought of as a far-off sci-fi hypothetical, but rather as an ongoing project which may well reach completion within the next decade or two (Posts #2–#3);
- In the absence of a good technical plan for avoiding accidents, researchers experimenting with brain-like AGI algorithms will probably accidentally create out-of-control AGIs, with catastrophic consequences up to and including human extinction (Posts #1, #3, #10, #11);
- Right now, we don't have any good technical plan for avoiding out-of-control AGI accidents (Posts #10–#14);
- Creating such a plan seems neither to be straightforward, nor to be a necessary step on the path to creating powerful brain-like AGIs—and therefore we shouldn't assume that such a plan will be created in the future "by default" (Post #3);
- There's a lot of work that we can do right now to help make progress towards such a plan (Posts #12–#15);
- There is funding available to do this work, including as a viable career option (Post #15).

For my part, I'm going to keep working on the various research directions in Section 15.2 above—follow me on Twitter or RSS, or check my website for updates. I hope you consider helping too, since I'm in *way* the hell over my head!

Thanks for reading, and again, the comments here are open to general discussion / ask-me-anything.