Connectomics seems great from an AI x-risk perspective

36

by **Steve Byrnes** 30th Apr 2023

Neuromorphic AI | Neuroscience | Whole Brain Emulation | AI | Frontpage

Context

Numerous people are in a position to accelerate certain areas within science or technology, whether by directing funds and resources, or by working in the area directly. But which areas are best to accelerate?

One possible consideration (among others) is the question: "Is accelerating this technology going to *increase* the chance that our future transition to superhuman artificial general intelligence (AGI) goes well? Or *decrease* it? Or make no difference?" My goal here is to try to answer that question for **connectomics** (the science & technology of mapping how neurons connect to each other in a brain).

This blog post is an attempt to contribute to Differential Technology Development (DTD) (part of the broader field of Differential Intellectual Progress°). Successful DTD involves trying to predict complicated and *deeply* uncertain future trajectories and scenarios. I think the best we can hope for is to do better than chance. But I'm optimistic that we can at least exceed that low bar.

My qualifications: I'm probably as qualified as anyone to discuss AI x-risk and how it relates to neuroscience. As for connectomics, I'm not too familiar with the techniques, but I'm quite familiar with how the results are used—in the past few years I have scrutinized probably hundreds of journal articles describing neural tracer measurements. (Think of neural tracer measurements as the traditional, "artisanal", small-scale version of connectomics.) I find such articles extremely useful; I would happily trade away 20 fMRI papers for one neural tracer paper. This post is very much "my opinions" as opposed to consensus, and I'm happy for further discussion.

TL;DR

- **Improved connectomics technology seems like it would be very helpful for the project of reverse-engineering circuitry in the hypothalamus and brainstem that implement the "innate drives" upstream of human motivations and morality. And that's a good thing!** We may wind up in a situation where future researchers face the problem of designing "innate drives" for an AI°; knowing how they work in humans would be helpful for various reasons°.
- **Improved connectomics technology seems like it would NOT be very helpful for the project of reverse-engineering the learning algorithms implemented by various parts of the brain, particularly the neocortex. And that's a good thing too!** I think that this reverse-engineering effort would lead directly to knowledge of how to build superhuman AGI, whereas I would like us to collectively make much more progress on AGI safety & alignment *first*, and to learn exactly how to build AGI *second*.
- **Improved connectomics technology might open up a path to achieving Whole Brain Emulation (WBE) earlier than non-WBE AGI. And that's a good thing too!** Generally, a WBE-first future seems difficult to pull off, because (I claim) as soon as we understand the brain well enough for WBE, then we already understand the brain well enough to make non-WBE AGI, and someone will probably do that first. But if we *could* pull it off, it would potentially be very useful for a safe transition to AGI. I have previously been very skeptical that WBE is a possibility at all, but when I imagine a scenario where radically improved human connectomics technology is available in the near future, then it does actually seem like a possibility to have WBE come before non-WBE AGI, at least by a year or two, given enough effort and luck.

1. Background considerations

1.1 The race between reverse-engineering the cortex versus reverse-engineering the hypothalamus & brainstem

My theory is° that parts of the brain (esp. cortex, thalamus, striatum, and cerebellum) are running large-scale learning algorithms, while other parts of the brain (esp. hypothalamus and brainstem) are doing "other stuff".

- The large-scale learning algorithms, I claim, are more-or-less the "secret sauce" of human intelligence. Once we understand these algorithms, people will almost immediately start building superhuman AGI. (Some nuances here°.)

- The "other stuff" contains, among other things, "innate drives", including drives that make us feel that being-in-pain is bad and eating-when-hungry is good, as opposed to the other way around. Importantly, some of these "innate drives" constitute the suite of human social instincts, which in turn are upstream of human morality, friendship, and so on (I claim).

I strongly believe that, other things equal, it would be good to reverse-engineer the hypothalamus & brainstem (and particularly how they lead to human social instincts) *first*, and reverse-engineer the cortex and other large-scale learning algorithms *second* (ideally with a very long gap between them). If we do it in the opposite order, we will wind up in a place where people are messing around with AGI-capable algorithms, and maybe they'll *want* to make these algorithms feel a drive to be compassionate etc., but they won't know how. Fuller explanation here° and here°.

Unfortunately, the status quo in neuroscience is the exact opposite. Almost 100% of the best researchers with a knack for AI and algorithms are focused on the cortex, striatum, and cerebellum. Those researchers seem to care about the hypothalamus & brainstem only in the tiny spots where it's directly interfacing with the learning algorithms—VTA, SNc, inferior olive, etc. In other words, to the extent that the brain is doing reinforcement learning, those researchers mainly care about "how do reward signals update the trained models", whereas I mainly care about "how are the reward signals calculated in the first place"? Making a bad situation worse, I think that understanding the cortex is objectively easier than understanding the hypothalamus and brainstem—compare cortical uniformity° on the one hand, versus my hypothalamus overview° on the other hand.

1.2 The race between reverse-engineering the cortex versus wall-clock time

Should we hope for "reverse-engineering the cortex" to happen earlier or later, in calendar time?

I see three main considerations in play.

(1) If AGI is invented later, we'll have more time for alignment research, and I think this is very good and very important—further discussion here°. So this consideration pushes strongly towards delaying reverse-engineering the cortex as long as possible.

(2) If the cortex is reverse-engineered later, it becomes more likely that we'll get AGI via a different independent AI research program. However, this possibility does not push me much

in either direction:

- I think that those "different independent AI research programs" are probably going to plateau, and we're going to get cortex-like learning algorithms regardless (see here°).
- If not—*if it's technologically possible at all* for large language models (LLMs) & related systems to scale to transformative superhuman AGI—then I expect that to happen so soon that no other research program will matter.
- *Even if* I thought that LLMs were not bound to plateau, *and* that Differential Technology Development could meaningfully shift the chances for cortex-like learning algorithms to beat LLMs to the superhuman-AGI finish line, I *still* wouldn't really be rooting for one side or the other—it seems pretty unclear to me which of those would be better or worse.

(3) Many other aspects of the world will change over time. The geopolitical and commercial competition landscape could get better or worse. Computer security could get better or worse. Chips will presumably get better and cheaper, unless the Taiwan fabs get destroyed in WWIII or something. Even though I expect LLMs and their successors to plateau before becoming capable enough to radically transform the world, they will nevertheless continue to improve and become widely used. With each passing year there's a chance of bioengineered super-plagues. Etc. I'm happy to discuss any of these in detail, but to sum up, none of them are pushing me strongly in one direction over another, certainly not compared to **(1)** above.

In summary, consideration **(1)** is the winner, and therefore I'm hoping for reverse-engineering the cortex to happen later rather than sooner.

1.3 The race between Whole Brain Emulation (WBE) versus other forms of AGI

Radical advances in connectomics would presumably be essential for WBE, so let's talk about that.

My sense is that AI x-risk people (including me) have traditionally treated the possibility of getting WBE before AGI as *desirable but unrealistic*. (For example, see the report from the 2011 "Singularity Summit".)

Why does WBE seem desirable? Because with a WBE, we kinda know what we're getting:

- A prosocial person will (supposedly—see below) become a prosocial WBE, whereas an AGI may have alien motivations;

- A reasonably smart person will become a reasonably smart WBE, whereas an AGI may have unknown and potentially superhuman capabilities.
- Etc.

By the way, if we have WBEs, what do we do with them? Possibilities include:

- We could make lots of WBEs, speed them up, and have them spend many subjective centuries in cushy VR environments pondering the problem of AI alignment and deployment, and hopefully they'd come up with better plans than we could;
- We could legalize WBEs and ban any other form of AGI forever (in principle);
- We could take WBEs to be the starting point for more powerful AGIs, e.g. keep their innate drives the same but massively increase the number of cortical columns or whatever. This isn't guaranteed to be possible or wise or safe, but maybe it is, I dunno.

Why does WBE seem unrealistic? It seems that, in the course of trying to do WBE, we would necessarily wind up understanding brain learning algorithms well enough to build *non*-WBE brain-like AGI, and then presumably somebody would do so before WBE was ready to go. For further discussion see my short post: Randal Koene on brain understanding before whole brain emulation°.

How might WBE become more realistic? In principle, if we have a giant perfectly-secret WBE project and infinite time until any other form of AGI, then sure, WBE-first is possible.

Leaving that aside, if there was a radical advance in connectomics, such that every neuron and synapse in a human brain could be mapped in the near future (i.e. *before* understanding the brain well enough for brain-like AGI), that's at least a start. Then, one might think, as soon as we understand the brain well enough for brain-like AGI, we're also *immediately*[1] ready to do WBE.

In other words, if we have complete human connectomes on a hard drive before anything else happens, then WBE wouldn't *win* the race against non-WBE brain-like AGI, but at least it might be close to a tie, and then *maybe* with very good secrecy (and luck[1]) we could hope for a few years of exclusively WBE.

A major complication is that a connectome is not enough for WBE. The problem is there are a lot of things not captured by the connectome, like glial cells, information storage in gene expression, extrasynaptic flows, neuropeptide° expression and receptors, etc. Or as I've written before, Building brain-inspired AGI is infinitely easier than understanding the brain°.

It's hopeless to measure and simulate everything down to the subcellular level—we need to disentangle what's essential, versus an implementation detail. But there's no way to do that without *understanding*. And how do we get that? I propose:

Understanding a part of the brain well enough to simulate it requires:

- (A) Constraints derived from brain observations (e.g. connectomic data, lesion studies, neural recordings, understanding of how neurons work, etc.);
- (B) AND "normative" ideas about what that part of the brain is designed to do;
- (C) AND algorithmic understanding of how (B) might plausibly get done;
- (D) AND enough person-hours to fit the previous three puzzle-pieces together.

(Related: Dileep George's "triangulation strategy")

As a consequence:

- If we *fail* in the Section 1.1 goal—i.e., if we come to understand the cortex before we understand prosociality-relevant parts of hypothalamus & brainstem—and try to make WBEs anyway, then they would have "the wrong innate drives" in ways that really matter. They might still be as smart as the source human, but they would be missing what makes them safe, and indeed, arguably what makes them human. Thus we lose *most* of the justification for wanting a WBE option in the first place. Maybe not all of it though! It might still be nice to have an option of running these weird, mad WBEs. Like, maybe we can run them for a minute or so—a short enough time that they will have not yet turned completely deranged—and ask them "what feels different and wrong?", and use that as part of how we iterate and get the right innate drives. Or something, I dunno.
- If we *succeed* in the Section 1.1 goal—i.e., if we come to understand the important innate drives in the hypothalamus & brainstem before we fully understand the cortex—then things are looking much better for *both* brain-like AGI *and* WBE. Nevertheless, while both are potentially good, I think there's still reason to prefer WBE over brain-like AGI in this scenario—particularly the fact that WBE would avoid the need for training-from-scratch, which in turn is a giant headache for safety°.

So in summary, the Section 1.1 race matters a whole lot, but regardless of how that turns out, I think it would be nice to have WBE on the table as an option. And a radical advance in connectomics would make that more plausible than I had been previously imagining.

2. Finally, that brings us to connectomics

I'm for it! 👍👍

First, **I see future advances in connectomics as *only slightly* helpful for better understanding the cortex learning algorithm.** (Recall from Section 1.2 above that the previous sentence is a ***good*** thing from my perspective!)

Why? In ML terms, if we measure tons of neuron connections in the cortex, we're mainly measuring *trained model* parameters, and only indirectly getting information about how the learning algorithm built that trained model in the first place.

(Recall that we already have probably hundreds of thousands of papers about how the cortex is structured, plus a connectome of 1mm³ of mouse cortex, and a human version coming soon. Here I'm merely saying that further radical advances in connectomics probably wouldn't make much difference for understanding the cortex learning algorithm *on the margin*.)

By contrast, **I see future advances in connectomics as *super*-helpful for better understanding the hypothalamus & brainstem.** See discussion here°—a lot of things in the hypothalamus & brainstem are hundreds of specific idiosyncratic circuits doing specific things in specific ways. Researchers mostly can't be bothered to measure all these things, except in special cases where it's relevant to well-studied diseases (e.g. there is abundant data on hunger-related hypothalamus circuits because of their relevance to obesity). Making connectomic data more cheap & abundant would help. It's not *sufficient*, as discussed in Section 1.3—we still need lots of interpretive labor and presumably follow-up experiments—but it's extremely helpful.

(If the hypothalamus & brainstem connectomic data can be connected to neuropeptide expression and receptors, then so much the better. If the connectomic data is in humans, better yet—I have serious concerns that human sociality-related innate drives may be importantly different even from chimps.)

And obviously, as discussed in Section 1.3, **radical advances in *human* connectomics would at least potentially put WBE on the table**, which also seems good.

3. Closing promo

I'm far from an expert on the state of connectomics technology, but after talking to someone in the field, I have a vague impression that I am not *crazy* to hope for whole primate-brain connectomes in the 2020s and whole human-brain connectomes in the 2030s, if all goes well. That's soon enough to be plausibly relevant for AGI, on my models.

One group that I happen to be familiar with is **E11 bio** in Alameda, CA, USA. My (arms-length) impression is that they're thoughtful, ambitious, in a hurry to move fast, and have a really cool technological approach. And they told me that their technology is supposed to be compatible with simultaneously measuring not only neuron paths, but also other molecular markers like neuropeptide receptors (see Section 1.3 above). As of this writing, E11 is hiring, including for non-bio roles like ML, and they're fundraising as well.

I don't mean to be partisan—if E11 has any existing or future competitors, great! I am rooting for them too.

(*Thanks Anders Sandberg, Adam Marblestone, and Justis Mills for critical comments on a draft; see also Adam's 2021 post along similar lines.*)

---

1. ^ Unfortunately, I can think of various possible reasons that, *even with a complete human connectome already sitting on our hard drive*, WBE *might* still require years more work after we know how to make brain-like AGI. **(1)** Maybe once we understand brain algorithms, we'll notice that the natural way to implement them on a chip is quite different from the natural way to implement them on biological neurons. That's no problem for brain-like AGI, but for WBE would require translating the human's existing memories into a different low-level format, which might not be possible losslessly. **(2)** Relatedly, maybe there will be a substantial period where we understand *most* pieces of the puzzle of how brain algorithms work, and then we can get brain-like AGI by using fully-artificial components for the missing pieces, but wouldn't yet be able to do WBE. **(3)** Maybe once we understand how brain algorithms work, we'll notice that there is additional non-connectomic data that we need for WBE, like the theory that animal memories are mostly encoded via within-neuron gene expression. (I happen to think this theory is wrong, but a much weaker version of it seems possible to me.) **(4)** An adult human has a "trained model" optimized to control a real brain and body connected to each other in the real world, with all of its warts and nuances, and this model may be maladapted to a new and very different "environment" and suite of actions. Body-brain interactions can be complicated. I guess people can adapt to things like gland removal and quadriplegia, so maybe this isn't a huge deal, but could still add time and complication to a WBE approach.

Neuromorphic AI 2 | Neuroscience 2 | Whole Brain Emulation 2 | AI 2 | Frontpage

Mentioned in

46 Four visions of Transformative AI success

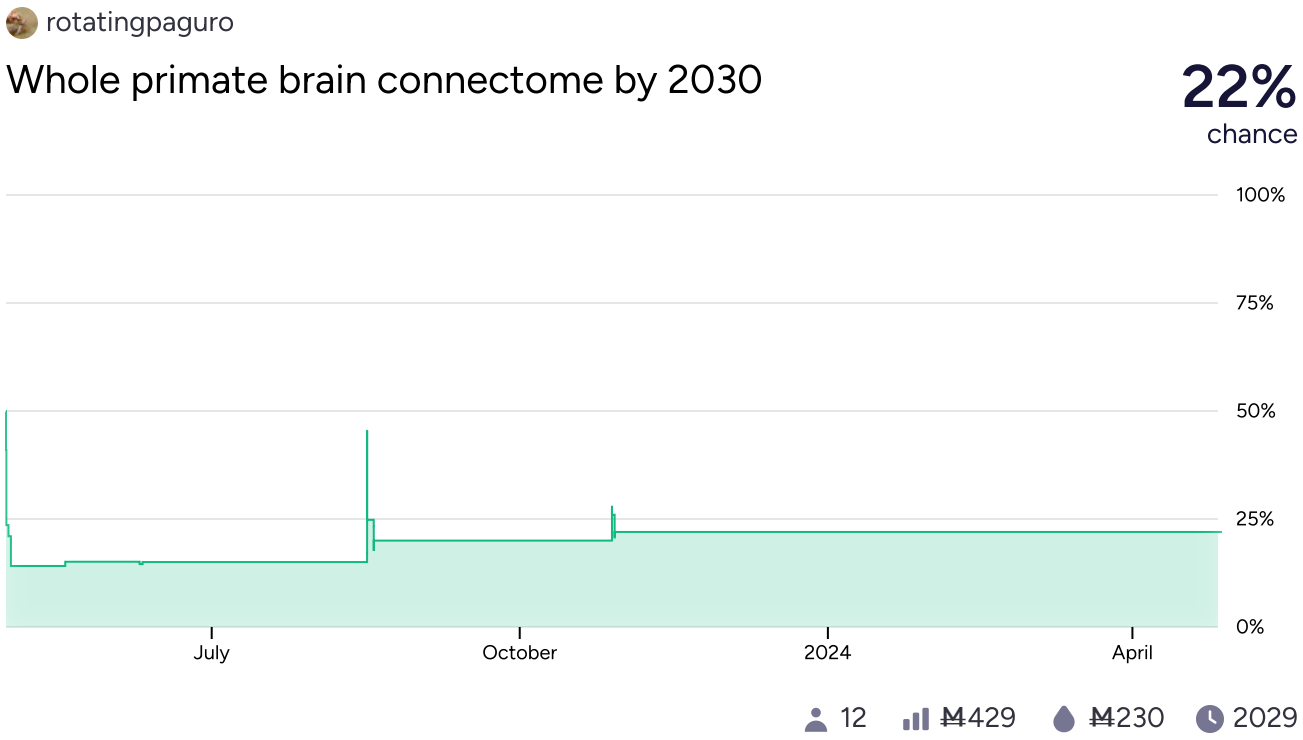
6 comments, sorted by top scoring

[−] **rotatingpaguro** 1y 4 1

> I have a vague impression that I am not crazy to hope for whole primate-brain connectomes in the 2020s and whole human-brain connectomes in the 2030s, if all goes well.

After reading the post "Whole Brain Emulation: No Progress on C. elegans After 10 Years" ° I was left with the general impression that this stuff is very difficult; but I don't know the details, and that post talks about simulation given a connectome, not getting a connectome, which maybe then is easier even for a huge primate brain, I guess? And I don't know what probability you mean with "not crazy".

A market here is thus apposite:



[−] **Steve Byrnes** 1y 4 2

> I don't know what probability you mean with "not crazy"

Me neither. I'm not close enough to the technical details to know. I did run that particular sentence by a guy who's much more involved in the field before I published, and he said it was a good sentence, but only because "not crazy to hope for X" is a pretty weak claim.

> After reading the post "Whole Brain Emulation: No Progress on C. elegans After 10 Years" ° ...

Yeah, the C. elegans connectome has been known for a very long time. The thing that's hard for C. elegans is going from the connectome to WBE. As crazy as it sounds, I think that there are ways in which a human WBE is *easier* than a C. elegans WBE. I talk about that to some extent here °.

[−] **Bogdan Ionut Cirstea** 1y 3 0

Related - I'd be excited to see connectome studies on how mice are mechanistically capable of empathy; this (+ computational models) seems like it should be in the window of feasibility given e.g. Towards a Foundation Model of the Mouse Visual Cortex: 'We applied the foundation model to the MICrONS dataset: a study of the brain that integrates structure with function at unprecedented scale, containing nanometer-scale morphology, connectivity with >500,000,000 synapses, and function of >70,000 neurons within a ~ $1mm^3$ volume spanning multiple areas of the mouse visual cortex. This accurate functional model of the MICrONS data opens the possibility for a systematic characterization of the relationship between circuit structure and function.'

The computational part could take inspiration from the large amounts of related work modelling other brain areas (using Deep Learning!), e.g. for a survey/research agenda: The neuroconnectionist research programme.

[−] **Vladimir Nesov** 1y 3 0

> Generally, a WBE-first future seems difficult to pull off, because (I claim) as soon as we understand the brain well enough for WBE, then we already understand the brain well enough to make non-WBE AGI, and someone will probably do that first. But if we could pull it off, it would potentially be very useful for a safe transition to AGI.

One of the dangers in transition to AGI, besides first AGIs being catastrophically misaligned, is first (aligned) AGIs inventing/deploying novel catastrophically misaligned AGIs, in the absence of sufficiently high intelligence to spontaneously set up effective security measures that prevent that. A significant jump in capabilities that doesn't originate from AGIs themselves doing work is safer in this respect, things like scaling of models/training that doesn't involve generating novel agent designs or mesa-optimizers. WBEs don't have that by default, even if they look much better on alignment.

[−] **rmorey** 9mo 2 0

Hi ! Very interesting post, I agree that connectomics might be most interesting for the "other stuff" and perhaps not the cortical areas. Well sourced post overall, but I wanted to add my perspective, as someone recently come to work in micro scale connectomics. I think you are not crazy to hope for those timelines, but I think you are unfortunately wrong. Led by our lab at Princeton, we have released the first adult whole brain fruit fly connectome: https://www.biorxiv.org/content/10.1101/2023.06.27.546656v1 it is a great milestone, and some are already starting to do whole drosophila brain emulation (https://www.biorxiv.org/content/10.1101/2023.05.02.539144v1) based on it, to some initial success. I think you should recalibrate your expectations, my understanding of the consensus, based on talking to my

colleagues (who are actual neuroscientists, unlike me, a software engineer) it's possible that we could make it to a whole mouse brain connectome in roughly a decade, but that is somewhat optimistic. There are fundamental barriers to scaling our imaging techniques that will need to be solved before scaling to larger brains. Additionally the greatest bottleneck currently is proofreading of the connectome. The automated processes will get better, but as of yet, human proofreading is still absolutely necessary, and obtaining ground truth for training better reconstruction models is likewise difficult. Overall, optimistically, I think whole primate brain connectomes are feasible in the next few decades. I am pretty firmly of the belief that AGI by other means will arrive well before WBE.

[−] **Steve Byrnes** 9mo 2 0

I'm not sure what you think my expectations are. I wrote "I am not *crazy* to hope for whole primate-brain connectomes in the 2020s and whole human-brain connectomes in the 2030s, if all goes well." That's not the same as saying "I *expect* those things"; it's more like "those things are not completely impossible". I'm not an expert but my current understanding is (1) you're right that existing tech doesn't scale well enough (absent insane investment of resources), (2) it's not impossible that near-future tech could scale much better than current tech. I'm particularly thinking of the neuron-barcoding technique that E11 is trying to develop, which would (if I understand correctly) make registration of neurons between different slices easy and automatic and essentially perfect. Again, I'm not an expert, and you can correct me. I appreciate your comment.

Crossposted to the EA Forum. Click to view.

Moderation Log