IFP

Mapping the Brain for Alignment

How to map the mammalian brain's connectome to solve fundamental problems in neuroscience, psychology, and AI robustness

Adam Marblestone and Andrew Payne

Mapping the Brain for Alignment

How to map the mammalian brain's connectome to solve fundamental problems in neuroscience, psychology, and AI robustness

Adam Marblestone and Andrew Payne

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

Frontier AI systems are neural networks that learn from a type of feedback called "reward signals" to interact better with humans. Yet, no scientist on Earth has a wiring diagram of the biological reward circuits that shape mammal brains to be (mostly) cooperative and resilient to stress. Until recently, creating a full brain wiring diagram with labeled cells and synapses, called a "connectome," for an animal like a mouse was theoretically a multi-billion, decade-long project. Today, new microscopes and data processing technology are pushing the price to tens of millions per brain, and compressing the duration of one of these mapping projects to potentially fit within a single congressional term.

We propose the Comparative Brain Connectome Initiative (CBCI): a NSF-supported Focused Research Organization that will execute a five-year, $150 million national effort to publish open-access, cell-level connectomes of five small-mammal brains chosen for contrasting social instincts. The deliverables — including imaging hardware, annotated circuit diagrams, cloud datasets, as well as a lean organization that can operate a cutting-edge neuro-imaging facility — create a template for embedding pro-social circuits in frontier AI models, accelerate drug discovery for circuit diseases, and lock in US leadership across fields as diverse as AI alignment, brain health research, and hardware development. Congress could provide dedicated funding for this initiative by providing an increase in

appropriations for the National Science Foundation by $30 million per year for 5 years.

Motivation

Frontier AI systems already rival humans at many economically critical tasks, yet our understanding of why they choose one action over another is remarkably hazy. Inspired by many of the features of biological brain structure and development, artificial neural networks are currently sculpted by reward learning through trial and error. Like animals, AIs can drift toward unhealthy behaviors because of reward signals.

Artificial intelligences, however, lack the layered, redundant circuitry that we have to keep impulses and drives in check. Mammalian brains developed over evolutionary time to include a network of structures deeper in the brain which issue multiple, sometimes competing, reward signals. These subcortical structures drive mammals towards behaviors like bonding, parenting, play, guiding many species' behavior towards something we recognize as adaptive and broadly pro-social. They align the short-term behavior of an animal and its learning from that behavior with evolutionarily worthwhile long-term goals.

Remarkably, modern neuroscience cannot yet point to a single mammalian species whose full reward circuitry has been mapped. Even though electron-microscope pipelines have produced cubic-millimetre chunks of cortex, new and more robust technology is needed to scale connectomics to entire mammalian brains. Those fragments are scientifically dazzling but strategically insufficient to guide AI development because they omit the deep-brain structures most relevant to reward, self-control, and sociality. Such circuits could give us vital clues as to how to design reward functions that lead to AI systems with desirable properties.

Without that map, AI-safety researchers are forced to design "alignment" techniques in the dark, through a murky trial-and-error of their own. Reinforcement-learning from human feedback — a workaround used by today's large language models — relies on crowdsourced thumbs-up and thumbs-down signals. This system, along with others that train the model's outward behavior, seems to leave its internal drives and hidden goals untouched, so the risk of

internal misalignment remains high. Consider, for example, the case of the sociopath who can masquerade as cooperative, but has no real empathy or care for their fellow person.

An abundance of recent research has shown that AI models undergoing testing for alignment can and will engage in deception, sandbagging, blackmail, and other worrying behaviors. We believe that these models will become more sophisticated in their ability to both execute and obscure misaligned behaviors. Thankfully, we are still within the critical window in which new approaches can be developed and tested. In this context, a comparative atlas of mammalian steering circuits would supply a reference design for robust synthetic motivational systems, exactly the sort of biological prior that scientists in AI laboratories could use in their research.

Importantly, the need for full connectomic maps and the tools to analyze them is just as acute in medicine as it is in AI. Autism, schizophrenia, major depression, and binge-eating disorders, among others, can increasingly be understood as "connectopathies": illnesses in which the gross anatomy of the brain remains intact but the pattern of long-range wiring is subtly different, leading to meaningful behavioral differences. Effective circuit-level therapies will likely remain elusive until we can begin to evaluate the healthy wiring diagram in its entirety.

For decades, the obstacle for a project like this was cost, projected to be billions of dollars with conventional approaches; a whole mouse brain contains roughly five hundred cubic millimeters of tissue, and the conventional approach of imaging it slice-by-slice with electron microscopes requires overcoming massive unsolved sample processing challenges and hundreds of person years for manually proofreading the outputs of computer vision neuron segmentation models. In the past eighteen months, technologies like rapid-scan light microscopes, thick-tissue nanoscale resolution tissue expansion chemistry, protein barcoding, and self-supervised segmentation networks have changed the equation. A connectome that was once projected to cost a billion dollars could in the near future be produced for nearer to $30 million, and the marginal price of a second or third connectomic mapping falls even further because the same robotic imaging-and-analysis line used for the first can be readily reused.

When the BRAIN Initiative was announced in 2013, participating agencies included NIH, NSF, DARPA, and IARPA. However, although Congress provided multi-year

funding for the NIH portion of the BRAIN Initiative in the 21st Century Cures Act, it did not provide support for NSF and other agencies. Given the importance of brain research not only for human health but for US leadership in AI, we think there is a strong case to provide an increase in the NSF budget for this initiative. The private sector is unlikely to step in because a fully open connectome is a public good; biotech and pharmaceutical firms are not currently focused on neural circuit mapping, nor can any one of them capture most of the upside if they were to. For their part, academic laboratories cannot direct dozens of engineers towards a single five-year goal without running afoul of tenure clocks and grant cycles. This is why a dedicated, mission-limited organization funded as public infrastructure is the only realistic path forward.

The economic stakes of doing this work are enormous: many estimates anticipate the deployment of AI agents on complex tasks with long durations in the coming years. In that horizon, the United States must possess a biological blueprint for safe reward design, so that these agents do not exhibit deeply uncooperative and antisocial behaviors as they traverse the open web and interact more and more with the physical world and its infrastructure. Public health experts expect mental illness to be the dominant cause of disability in the same time frame. Providing a true circuit-level map of the mammalian brain provides researchers across all these fields with a much better starting point for their research.

Solution

The scientific problems that shaped progress for the past century — sequencing a genome, mapping the cosmos, eradicating a virus — have always required operating in the right institutional form in parallel with conducting the right experiments. Twenty years ago the Human Genome Project succeeded not because biologists suddenly became more insightful, but because they were embedded in a consortium that set clear milestones, acquired industrial-scale sequencers, and released the results to the field for which they were laying the foundation. In the same spirit, today's hardest technical bottlenecks — whole-brain connectomes among them — demand a model that fits them well, taking the best elements from academic collaborations, industry start-ups, and government-funded project-based research programs.

The Focused Research Organization (FRO) fills that gap. Pioneered by Convergent Research (which was co-founded by one of the authors of this proposal), this form is already seeing adoption across the international science ecosystem. A FRO is a time-limited, mission-driven, non-profit startup that hires a full-time interdisciplinary team, budgets enough capital to operate a full engineering pipeline, and then transitions off its initial funding once it has built its target product and distributed it to an ecosystem of users. The approach is proving itself in the technology-development program of one of Convergent Research's FROs called E11 Bio, (helmed by the other author of this proposal) which is working to reduce the cost of optical connectomics by an order of magnitude. The technology is ready for a sibling FRO to apply these techniques and run them continuously until five mammalian brains are fully mapped and openly shared.

Congress should authorize and fund the Comparative Brain Connectome Initiative (CBCI) as a five-year FRO operating under the National Science Foundation's existing Other Transaction Authority. The FRO model is purpose-built for this sort of public good: it hires a full-time, interdisciplinary team, nests engineering and biology under one roof, issues milestone-driven sub-awards, and then winds down at sunset, handing its equipment to successor projects.

CBCI would establish a high-throughput optical-connectomics pipeline that can process hundreds of thick tissue slabs, label them with multiplexed molecular barcodes, image them volumetrically at the nanoscale, and detect neuron shapes and synapses. A parallel cloud-hosted AI-segmentation cluster could convert raw imagery into 3D reconstructions and ultimately a navigable wiring diagram, while an electron-microscope node would provide ground-truth verification. Because the capital equipment is a dominant expense, imaging additional brains becomes much cheaper once the pipeline is running. Therefore, the program can deliberately commit to five species: male and female laboratory mice; the prairie vole, whose monogamous bonding offers an extreme contrast; the tree shrew, closer to primates and highly territorial; and the naked mole-rat, a eusocial outlier whose unusual cooperative behavior may reveal alternate solutions to reward design. A comparative set of five connectomes exposes conserved versus divergent circuit motifs and prevents researchers from over-fitting findings to a single animal.

CBCI aligns with the "X02 (Execution) Awards" of the X-Labs proposal submitted by Caleb Watney in this same collection. In this proposal, Execution Awards are a

format for awarding funds to startup-like research organizations that are smaller, faster-moving, and less encumbered by bureaucratic or political inertia. They aim to support high-risk, high-reward basic science by consolidating administrative overhead within research institutions, thereby freeing scientists from excessive grant writing and reporting burdens.

The program's formal milestones are straightforward. In the first year CBCI, purchases and installs microscopes, validates sample preparation, and releases a ten cubic-millimeter scale pilot optical connectomic image volume so the broader community can accelerate building analytic tools. By the end of year two the facility delivers a fully segmented male-mouse connectome annotated for all major neurotransmitters and cell types. Year three focuses on reward-circuit labeling and on shipping a cloud-native viewer that any AI lab can query. Year four adds three further species and runs a public "motif discovery" hackathon that invites alignment researchers to search for steering sub-circuits. Year five finalises datasets, opens the facility to cost-neutral external users, and convenes a joint workshop with DARPA, IARPA, NIH, and other agencies to plan therapeutic and brain-inspired neuromorphic spin-offs.

The total federal cost comes out to $150 million: roughly $50 million in capital for microscopes, robotics, and storage arrays; and a $100 million in operating expenses spread over personnel, consumables, cloud compute, and data-commons maintenance.

This work supports AI safety but has reverberations far beyond it. Neuropsychiatric teams can gain an atlas for circuit-level drug discovery, particularly in the hypothalamus (already a target for GLP-1 agonists in obesity). DOE researchers could obtain a concrete wiring template for energy-efficient neuromorphic chips at a moment when data center power budgets are becoming an enormous issue. Graduate students across the country would receive training in large-scale AI-driven biological data analysis, neural network reverse engineering and brain-inspired AI engineering, skill sets now in chronic shortage across the United States. And the broader public, whose tax revenue drives the effort, receives a user-friendly portal not unlike the UCSC Genome Browser (a type of Google Maps for DNA), giving citizen scientists and students real access to the wiring of the social brain.

Further resources

- Wellcome Trust, "Scaling up connectomics: the road to a whole-mouse-brain connectome," 2023.

  Wellcome Trust white-paper.

- NIH Advisory Committee, "A Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative 2025 Roadmap," 2014.

  U.S. NIH Advisory Committee report.

- Caleb Watney, "Launching X-Labs for Transformative Science Funding," n.d.
- David Markowitz, "Solving intelligence requires new research and funding models," The Transmitter, n.d.
- Steven J. Byrnes, "Intro to brain-like-AGI safety," 2022.
- Larry F. Abbott et al., "The mind of a mouse," Cell 182, no. 6 (2020): 1372–1376.
- Patrick Mineault et al., "NeuroAI for AI safety," arXiv preprint, 2024.
- A. Paul Alivisatos et al., "A National Network of Neurotechnology Centers for the BRAIN Initiative," Neuron 88, no. 3 (2015): 445–48.