Foresight Institute 2024 Workshop

Foresight Neurotech, BCI and WBE for Safe AI Workshop

LIGHTHAVEN, BERKELEY, CA, USA

21 & 22 May, 2024

Report writer: Naveen Rao

Table of Contents

Foreword

Whole Brain Emulation (WBE) represents a promising technology frontier for creating human-aligned software intelligence based on advances in our understanding of human biology. Recently, there has been a shift to update AGI timelines towards earlier development, raising safety concerns. This has led to considerations whether Brain-Computer Interface (BCI) or WBE approaches to Neurotechnology could be significantly sped up, producing a differential technology development re-ordering that may reduce the risk of unaligned AGI by the presence of aligned software intelligence.

Our 2023 workshop in this series, chaired by Anders Sandberg of the University of Oxford, explored AGI and WBE timelines, assessed which WBE approaches could be accelerated through coordinated efforts, and evaluated potential risks associated with these technologies. Building on this foundation, this year's event focused on emerging advances in AI, neuroscience, neural engineering, and other areas. This work complements our ongoing Fellowship Program and AI Safety Grant, which fund various WBE projects for safe AI.

Held over two days in Berkeley, California, this workshop brought together over sixty researchers, entrepreneurs, and funders working on neurotechnologies, particularly WBE approaches that could be safely and expediently achieved within relatively short AGI timelines. The event featured introductory keynote presentations followed by working groups exploring highlighted challenges and promising focus areas.

This report contains summaries and recordings of the presentations and ensuing project collaborations. You can access the corresponding presentations by clicking on the play icons in the images.

We extend our heartfelt gratitude to all participants for their work and collaboration. A special thank you also goes to our sponsors, AE Studio and Protocol Labs, for subsidizing the attendance of junior researchers. Your support was crucial for the success of this workshop.

We look forward to next year's workshop to review and build on projects initiated in 2024. If you are interested in advancing this area as a researcher, practitioner, or funder, we welcome you to reach out.

Best regards,

Allison Duettmann
CEO, FORESIGHT INSTITUTE

a@foresight.org

Participants

Adam Safron

JOHNS HOPKINS UNIVERSITY

Aidan O'Gara

CENTER FOR AI SAFETY

Alexandra Tuononen

STANFORD UNIVERSITY

Allison Duettmann

FORESIGHT INSTITUTE

Anders Sandberg

INSTITUTE FOR FUTURE STUDIES

Andrew Payne

E11 BIO

Anita Fowler

CARBONCOPIES FOUNDATION

Ariel Zeleznikow-Johnston

MONASH UNIVERSITY

Aurelia Song

NECTOME INC.

Benjamin Korpan

BOOTSTRAP BIO

Bobby Kasthuri

UNIVERSITY OF CHICAGO

Bogdan Ionut Cirstea

INDEPENDENT

Catalin Mitelut

NYU

Chase Denecke

BOOTSTRAP BIO

Claire Short

ATHENA

Daniel Berger

HARVARD UNIVERSITY

Participants

David McSharry
KREOH

Diogo de Lucena
AE STUDIO

Dmitri Mitya Chklovskii
SIMONS FOUNDATION

E.J. Chichilnisky
STANFORD MEDICINE

Frank Stegert
MAYAA

Garrett Baker
INDEPENDENT

Jacob Cannell
INDEPENDENT

Jan Kirchner
OPENAI

Janis Hesse
UC BERKELEY

Jed McCaleb
VAST

Joanne Peng
BOYDEN LAB

Johan Winnubst
E11 BIO

John Cumbers
SYNBIOBETA

Joseph Carlsmith
OPEN PHILANTHROPY

Juan Benet
PROTOCOL LABS

Judd Rosenblatt
AE STUDIO

Leonardo Christov-Moore
INSTITUTE FOR ADVANCED CONSCIOUSNESS STUDIES

Laria Reynolds/janus
CONJECTURE

Participants

Lina Paiz

STANDFORD UNIVERSITY

Lisa Thiergart

MIRI

Logan Collins

WASHINGTON UNIVERSITY

Marcus Kaiser

UNIVERSITY OF NOTTINGHAM

Maria Avramidou

UNIVERSITY OF OXFORD

Matias Serebrinksy

PSYMED VENTURES

Matthew McDougal

OKLAHOMA MEDICAL RESEARCH FOUNDATION

Max Hodak

SCIENCE

Michael Andregg

FATHOM RADIANT

Michael Skuhersky

MIT

Nathan Helm Burger

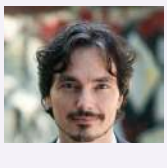
INDEPENDENT

Niccolò Zanichelli

UNIVERSITÀ DEGLI STUDI DI PARMA

Nikola Markov

THE BUCK INSTITUTE FOR RESEARCH ON AGING



Patrick Mineault

MILA

Philip Shiu

EON SYSTEMS

PK Douglas

NEUROTRUST AI

Randall A. Koene

CARBONCOPIES

Roman Bauer

UNIVERSITY OF SURREY

Participants

Saturnin Pugnet

WORLDCOIN

Sharena Rice

SANMAI TECHNOLOGIES PBC

Tiffany Yang

UNIVERSITY OF WASHINGTON

Todd Huffman

E11 BIO

Tracy Brandmeyer

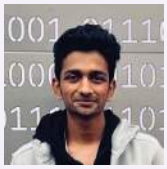
BRAINMIND

Tsvi Benson Tilsen

MIRI

Xiecheng Shao

USC



Viraj Chhajed

FORESIGHT PRODIGY FELLOW

Zan Huang

FORESIGHT FELLOW

Participant Group Photo

About Foresight Institute

Founded in 1986, Foresight Institute supports the beneficial development of high-impact technology to make great futures more likely. We focus on science and technology that is too early-stage or interdisciplinary for legacy institutions to support, such as longevity biotechnology, molecular machines, brain-computer interfaces, multipolar AI, or space exploration. We award prizes, offer grants, support fellows, and host conferences to accelerate progress toward flourishing futures and mitigate associated risks.

Workshop Chairs

Allison Duettmann

FORESIGHT INSTITUTE

Allison Duettmann is the CEO of Foresight Institute, directing programs in Intelligent Cooperation, Molecular Machines, Biotech & Health Extension, Neurotech, and Space. She founded Existentialhope.com, co-edited Superintelligence: Coordination & Strategy, co-authored Gaming the Future, and co-initiated The Longevity Prize. Duettmann advises companies such as Cosmica and serves on the Biomarker Consortium's Executive Committee. She holds an MS in Philosophy & Public Policy from the London School of Economics, focusing on AI Safety.

Anders Sandberg

INSTITUTE FOR FUTURE STUDIES

Anders Sandberg's research centers on estimating the capabilities and underlying science of future technologies, methods of reasoning about long-term futures, existential and global catastrophic risk, the search for extraterrestrial intelligence (SETI), as well as societal and ethical issues surrounding human enhancement. He coauthored the original WBE Roadmap.

Executive Summary

2024 may be remembered as the year that pursuit of AGI moved from the edges onto center stage. With the world's leading corporations shifting the global economy in pursuit of development of ever-more powerful AI, calls for responsibility and safety are at risk of becoming footnotes in the race for breakthroughs.

The moment calls for critical thinking and agency. Rather than ask "if" AGI will be harmonized with humanity's interests and developed in ways that safeguard the privacy, safety, health, and well-being of people around the world, as leaders we must ask "how" and roll up our sleeves.

WBE represents a promising technology frontier for creating or aiding in the creation of human-aligned software intelligence. Proponents of aligned AGI need more than ethical guidelines and safety frameworks. They require solutions-oriented thinking, rooted in technical expertise and the best available context. Only then will actionable solutions, built on viable timeframes and predictable financial scales, emerge into sight.

Foresight: A Catalyst for Action

In May 2024, 60 experts convened in Berkeley, California for a two-day workshop to explore the potential of WBE and neurotechnologies to guide alignment in AGI.

The workshop created a unique forum to share and cross-pollinate ideas, hold critical discussions, and establish common knowledge from which to build consensus. In a series of short plenary talks, global experts in neuroscience research, AI, neural engineering, and other subjects shared an overview of their work and perspective on the alignment challenge. Each session was followed by a brief Q&A discussion.

To guide discussions productively, Foresight Institute established multiple pathways towards alignment at the outset. Following three rounds of presentations, participants voted on various approaches and then self-assembled into breakout groups to explore pathways in greater depth. Participants reconvened to present their sessions, debate feasibility, risks, and challenges.

Prioritizing Neurotech Pathways Towards AI Safety

The following pages contain a summary of key sessions from the workshop. These sessions have been organized for easy navigation along four pathways: Hi-fi emulation, Lo-fi emulation, BCIs, and development of Prosocial agents.

The report also catalogs considerations to prioritize future research projects, rooted in technical discussions, real-world examples, and best-available estimates of developmental cost and timelines.

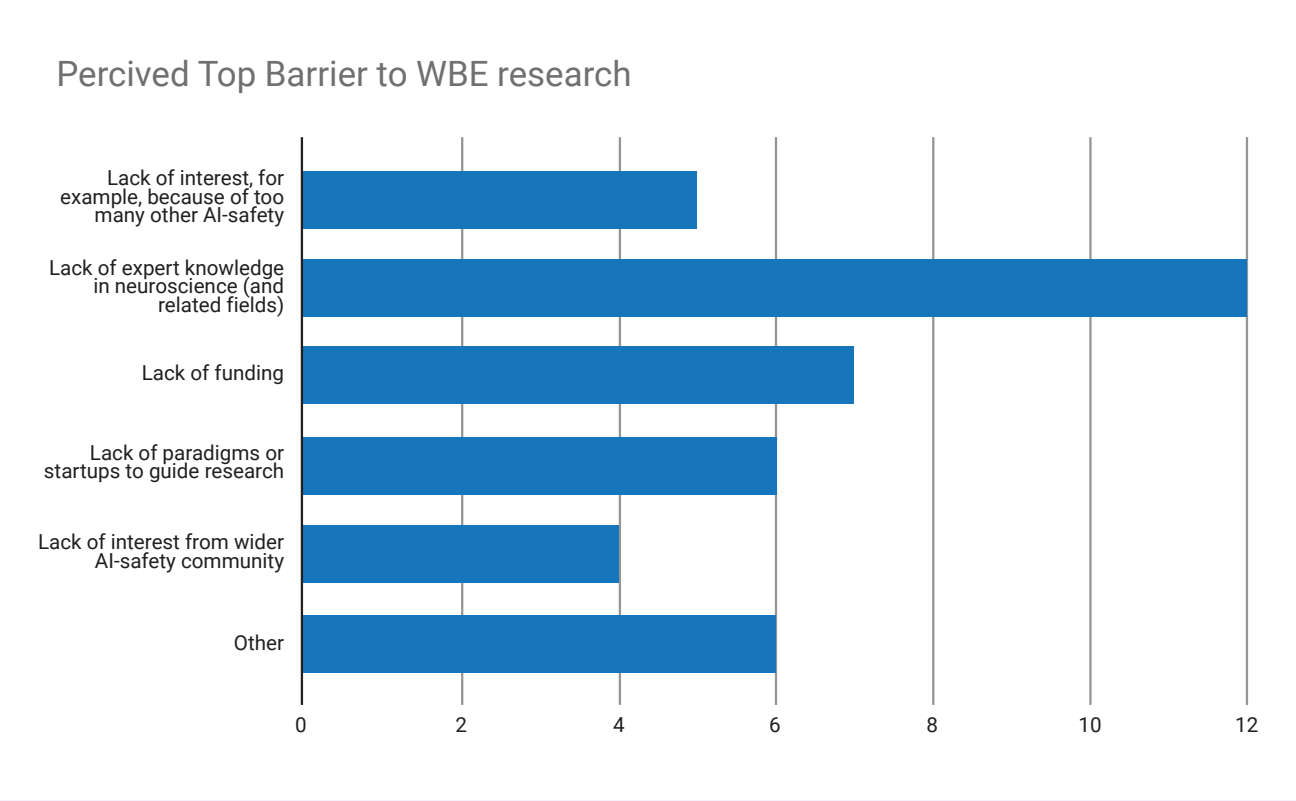
These pathways have been integrated into a single frame for comparison, establishing common parameters of time and resources, risks, benefits, and feasibility of impact. They are followed by proposals as well as examples of specific projects funded by the Foresight Institute's grant programs, described herein.

WBE for Aligned AGI

While AI has progressed dramatically over the last decade, 2024 may be remembered as the year that "AGI" became a feature, not a bug, of corporate R&D efforts. Calls for safety and concern have taken a backseat to competitive pursuit.

Interest in creating aligned AGI may not be sufficient without a clearer strategy for how alignment might be achieved. WBE is a concept that holds the potential to advance our understanding of human nature in such a way that enables alignment. In 2008, in a landmark WBE Roadmap report, Anders Sandberg defined WBE as "the possible future one-to-one modeling of the function of the human brain."

And yet, WBE remains a theoretical construct, as technology has only recently advanced to the point where such models are plausible. Key challenges to the successful emulation of the human brain include a lack of consensus on the best path forward, compounded by uncertainty about both the scale of funding and timelines required for each potential pathway.



A participant survey revealed perceived barriers to WBE Research

The workshop commenced in the spirit of establishing consensus on feasibility, defining clear success criteria, or prioritizing strategies to implement. It closed with ideas for how to build the next roadmap for WBE.

Naysayers are Your Unpaid Focus Group

Anders Sandberg, Institute for Future Studies

Sandberg discusses strategies for advancing the WBE community as a whole, via engaging with critics who can provide valuable insights to refine the community's approach. Firstly, he identifies several common arguments against WBE, including ethical concerns, technological limitations, and lack of clear definitions for success. To address these, he proposes developing clearer metrics and benchmarks, exploring technological solutions, and organizing workshops to discuss the ethics and consequences of future WBE. His conclusion is that, by treating naysayers as an unpaid focus group this community can strengthen its strategies and make meaningful progress.

An MVP of the WBE Challenge

Randall Koene, Carboncopies

Koene discusses the "translation challenge," which involves the translation of connectome data into functional models that can be run, an important missing link in accelerating neuroscience R&D. The goal of emulation is to reconstruct working neural circuits that process information and reproduce cognitive behavior, and the translation challenge is one step further from 3D reconstruction.

To address this challenge, Koene proposes generating ground truth data sets, synonymous with biological samples, to make measurable progress. A platform has been created to allow researchers to start with simple problems and progress to more complex ones. The success criteria for WBE, whilst numerous, have been condensed to eight, and the aim is to create end-to-end examples that demonstrate the possibility of WBE, validated with known ground truth systems. The ultimate goal is to create models specific to a particular animal that can reproduce the same memory, leveraging AI's ability to work with understandable data.

Towards the Next WBE Roadmap

- **Anders Sandberg**, Institute of Future Studies
- **Ariel Zeleznikow-Johnston**, Monash University
- **Lisa Thiergart**, MIRI
- **Nathan Helm Burger**, Independent
- **Randall A. Koene**, Carboncopies
- **Naveen Rao**, Neurotech Futures

This working group explores the WBE Roadmap Project, updating the 2008 Sandberg & Bostrom roadmap to reflect current brain emulation methodologies. Led by a two-person team to ensure completion, the project incorporates diverse expert input. The roadmap includes an introduction to WBE, chapters on success criteria and validation, main approaches and timelines, and extensive appendices covering topics from scanning technologies to AGI safety applications.

The group emphasizes the importance of defining clear emulation success criteria to guide research priorities. They estimate a few months for completion and suggest that dedicated oversight could be beneficial. Immediate steps include regular editorial meetings and phased review rounds. While acknowledging roadmaps' potential to aid resource allocation and accelerate progress, the team stresses the need for ongoing scrutiny and updates to avoid perspective lock-in. The project's relevance to AI safety is noted, highlighting WBE's potential to produce "aligned by default" entities with superhuman capabilities.

Pathways to Aligning AGI

The heart of the workshop involved generating and comparing promising approaches on impact, feasibility, benefits and risks across four primary pathways to alignment: Hi-fi, Lo-fi, BCI-based, and Prosocial.

- Hi-fi approaches generally refer to the creation of a digital human brain, developed fully, or near-fully, based on biological models and application of advanced software.
- Lo-fi approaches focus on developing whole brain models based on partial data or brain samples, animal models, and/or advanced technology.
- BCI - By augmenting human beings ability to use AI, next-generation biotechnologies offer possibilities to explore novel means of aligning AGI.
- Prosocial approaches focus on incorporating specific human traits and values into AI, to create more human-like and human-compatible AI systems.

The following summaries encapsulate selected sessions from the workshop. Summaries and comparisons of these approaches are compiled in a subsequent section of this report.

HI-FI APPROACHES

Advances in technological discovery in hardware and software are creating an emerging opportunity to create a digital representation of the human brain. With synergies between breakthroughs in brain mapping, neural networks, understanding neuronal structure and function, the possibility of creating a digital brain of high fidelity is becoming less theoretical.

Scaling up Ultrastructural Brain Mapping to Cubic Millimeters and Beyond: Lessons from our Recent Study on Human Cortex

Daniel Berger, Harvard University

Berger discusses their recent groundbreaking study, a collaboration between the Harvard Lyman Lab and Google, which presented a computationally intensive reconstruction of the ultrastructure of a cubic millimeter of human temporal cortex, producing 1.4 million gigabytes of data. The brain sample, acquired in 2014, was imaged within a year by lead author Alexander Sheno using a method developed by Kenneth Hayworth. Published in BioArchive in 2021 and recently published for Science, the team used Google's computational resources for image stitching, alignment, storage, and segmentation. He highlights how the team identified approximately 50,000 cells, with oligodendrocytes being the most numerous, and reconstructed around 150 million synapses. He notes that their pioneering work has gained significant media attention for its detailed imaging.

Mapping the Human Brain with Expansion X-Ray Microscopy

Logan Collins, University of Washington

Collins discusses combining expansion microscopy with x-ray microtomography to image the human brain, to overcome current obstacles limiting scaling from mouse brains to human ones. Via a Foresight AI Safety Grant, collaborating with Panluminate and Synchrotron facilities, he aims to develop this technology, potentially imaging the entire human connectome within a year of continuous imaging. Starting with small tissue samples, Collins and colleagues have written a preprint comparing prospects for whole-brain imaging technologies. Additionally, he discusses that he is seeking a grant to develop a spatially targeted method for growing brain parts using focused ultrasound and a blood-brain barrier crossing virus, initially targeting Alzheimer's treatment, but recognizing its potential for enhancing human intelligence.

Large-Scale High-Density Brain-Wide Neural Recording in Nonhuman Primates

Janis Hesse, UC Berkeley

Hesse stresses the importance of understanding consciousness for AI safety and preserving one's consciousness when uploading or emulating the mind in a machine. He explains how physical processes, such as electrical activity in microscopic neurons, can give rise to conscious experience, and how his team has developed neuropixels probes to study it. They test the analysis by synthesis framework, suggesting that conscious perception arises from the activation of the feedback branch of the loop, while unconscious perception stems from the feedforward branch.

The brain alternates between these waves, encoding conscious percepts during feedback and unconsciously representing incoming physical stimuli during feedforward. The study indicates that conscious perception may be discrete, stitched together from epochs in time, and only occurs during feedback waves when the brain generates predictions based on its world model. Hesse's team has not only read out conscious precepts but also written hallucinated precepts through electrical microstimulation and other means.

The Smart Neuron: A Biologically-Inspired Foundational Building Block of Neural Networks

Dmitri Mitya Chklovskii, Simons Foundation

Chklovskii discusses a normative neuroscience approach to simulate large-scale whole brain connectomes and understand neuronal function. He argues that current models of biological neurons are either too simplistic or overly detailed for WBE. Chklovskii's group, Normative in Neuroscience, aims to generate a computational primitive that can be placed into the nodes of the connectome, enabling whole brain simulation. He explains this approach as deriving a solution to an optimization problem as an algorithm that meets the constraints of biological hardware. The team then maps this algorithm onto a biological neuron, capturing essential physiological properties, such as membrane potential voltage dynamics and synaptic plasticity.

Towards a White Matter Ephaptic Coupling Model of 1/f Spectra

P.K. Douglas, Neurotrust AI

Douglas explores the potential of ephaptic coupling in EEG and white matter, arguing that a generative model with biologically plausible subcomponents, capable of recapitulating EEG data and performing tasks, would be more compelling than current paradigm-bound theories. An ideal model should explain various EEG phenomena, such as cross-frequency phase amplitude couplings and individual alpha peak differences. While classical models consider axonal conduction negligible due to asynchronous summation, Douglas suggests that white matter may contribute to EEG signals through ephaptic coupling, promoted by unmyelinated neurons increasing fiber packing density and transverse resistivity within white matter microstructure. The article presents a model using cable theory and human connectome project data to estimate the ephaptic coupling strength.

LO-FI APPROACHES

Low fidelity approaches seek to build pathways towards alignment that originate from existing foundational research and development in areas of connectomics, evolutionary biology, and artificial intelligence. Incorporating our emerging understanding of neural dynamics, research frontiers with animal models, novel virtual neuroscience methods, or other advances, proponents of low-fi pathways suggest a faster, more cost-effective strategy is best.

Scale-Dependent WBE: An Experimental Driven Research Program

Catalin Mitelut, NYU

Mitelut focuses on functional neural data recording and its applications in WBE. He aims to record rodent electrophysiology and behavior to build a neural emulator using language models, generating predictions of both behavior and neural states. Mitelut considers whether modeling neural dynamics alone could achieve WBE, bypassing extensive mechanistic neuroscience research. Particularly intrigued by the complexity of spontaneous neural dynamics and their role in the neuroscience of agency, he plans to conduct extensive electrophysiology recordings of rodents in open field environments. By capturing these intricate dynamics, Mitelut hopes to deepen our understanding of WBE and replicate the success of language models.

Connectomics and AI Safety

Andrew Payne, E11 Bio

Payne puts forward that using the expertise of neuroscientists studying human-aligned general intelligence could significantly accelerate progress towards AGI alignment. He identifies three promising areas within connectomics: brain-inspired AGI, which studies neural circuits encoding social instincts and learning control; WBE, which reduces the number of free parameters in AGI models; and BCIs, which can keep humans involved in exploring AGI alignment strategies. Whilst current electron microscopy tools have been successful in smaller-scale neuroscience studies, Payne emphasizes the need to scale them up to examine entire mouse brains. He suggests that these kinomics applications could yield results within the next decade as technical challenges are overcome, potentially playing a crucial role in accelerating progress towards AGI alignment.

Frontiers of Whole Brain Emulation

Aurelia Song, Nectome Inc.

Song proposes a new approach to WBE using foundation models, challenging the traditional bottom-up emulation method. They outline two potential pathways: an fMRI-only approach, which creates a video model based on brain imaging data, and a connectome pathway, which models neural connections. Song argues that combining these methods may deliver more accurate brain emulations. Whilst acknowledging current limitations in parsing human language with computer science tools, they highlight the success of foundation models in other domains, such as in weather prediction. Song concludes that a multi-faceted approach, leveraging various techniques, is most likely the most promising route to achieving effective WBE, potentially opening new avenues for advancing research and development.

From Foundation Models in Neuroscience to WBE

Patrick Mineault, Mila

Mineault is working on linking foundation models for the brain to lo-fi WBE to build virtual brains with AI and accelerate neuroscience. He underscores that virtual neuroscience allows for fully observable states, reproducibility, causal ablations, interventions, and less reliance on animals, making it faster and more efficient than traditional neuroscience – which often requires invasive procedures and slow knowledge creation and publication cycles.

Mineault estimates that there may be close to the right amounts of data for this type of training, with about a million hours of neural data corresponding to a trillion tokens. Despite technical issues such as comparing data from different animals, progress is being made through alignment and latent space mapping allowing for the comparison of neurons from different animals. With the potential for the vast acceleration of neuroscience, Mineault believes that virtual neuroscience should be pursued by neuroscientists.

BCI-BASED APPROACHES

While today's implanted neural devices represent an early effort at brain-computer interfaces, advances are emerging. In mid-2024, over half a dozen companies had registered their products with the FDA for pursuit of regulatory clearance, via the investigational device exemption (IDE) pathway.

One of those companies, Neuralink, has openly stated their long-term focus of human enhancement, while another, Synchron, recently announced a live integration with GPT-4.0 in one of their trial participants to respond to AI prompts via BCI control. Despite this pace of innovation, BCI and other in-vivo enhancement represent an expensive, time-consuming, and high-risk pathway when viewed through a lens of AGI alignment.

Catalyzing Innovation in BCI

Tracy Brandmeyer, Brainmind

Brandmeyer discusses her organization, Brain Mind, which is a multi-sector community of researchers, venture capitalists, philanthropists, and entrepreneurs to develop innovation in BCI. They combine philanthropic funds and venture capital to create self-sustaining companies that drive meaningful change. With a focus on brain mapping, they emphasize the development of hybrid funding models. Brandmeyer stresses Brain Mind's commitment to supporting collaboration among neuroscientists, policymakers, ethicists, and other stakeholders to create innovative funding structures that propel the advancement of neuroscience. Through these efforts, she argues that Brain Mind aims to unlock the potential of neurotechnology and its applications in the BCI space.

Bi-Directional Neural Interface for Vision Restoration, Science, and Augmentation

E.J. Chichilnisky, Stanford Medicine

Chichilnisky explores the potential of electronic retinal implants to treat or cure blindness caused by degenerative diseases that damage the retina, the neural tissue responsible for encoding visual information. The concept involves developing an electronic device that stimulates the remaining retinal ganglion cells, inducing them to send artificial visual signals to the brain, thereby reproducing

the natural neural code of the retina. However, Chichilnisky highlights that current retinal implant technology is ineffective, and extensive scientific research would be needed to better understand how to restore vision. He emphasizes the importance of understanding the diversity of retinal signals, as the retina contains various cell types that perform different visual processing tasks and transmit signals to distinct brain outputs. He concludes that comprehending this complexity is crucial for understanding retinal function and developing more effective vision restoration techniques.

PROSOCIAL APPROACHES

Incorporating humanity's values vis a vis emergent understanding of self-modeling, empathy, interpersonal psychology, and other pro-social attributes into AI represents another pathway to alignment. These approaches represent a lower-cost, faster approach to consider.

Alignment and Value Drift in WBE

Maria Avramidou, University of Oxford

Avramidou explores how WBE could help address the alignment problem, focusing on value drift. She argues that successful brain emulations should have the same values as biological humans, eliminating the need to specify initial values. Using evolutionary theory, she introduces the concepts of natural selection and replicators to understand how values and structures change over time, leading to questions about which entities are subject to natural selection, who benefits long-term, and who adapts as a result. Avramidou suggests that an entity's fitness determines long-term benefits from selection, and that behavior adaptation and value drift depend on the environment, which can be shaped for positive outcomes. She urges consideration of incentives emulations might employ, such as self-copying, self-sacrifice, and cooperation with humans, and how these can benefit humans as secondary beneficiaries.

Self Modeling in Neural Networks

Diogo Di Lucena, AE Studio

Di Lucena discusses how he and his collaborators implement self-modeling in neural networks, inspired by Graziano's attention schema theory, to reduce model complexity and foster pro-social behavior in multi-agent systems. By creating simplified models of one's own attention, the same mechanisms used to model others' attention, the team demonstrates that self-modeling leads to reduced complexity, making it easier for agents to model the system and potentially enhancing prosocial behavior. Evaluating their method on various tasks and models, the authors find consistent results in complexity reduction, suggesting that their approach, grounded in neuroscience, is distinct from other regularization methods and applicable across diverse contexts to improve model complexity and pro-social behavior in multi-agent settings.

Neurotechnology and the Weak to Strong AI Generalization Framework

Jan Kirchner, OpenAI

Kirchner calls attention to the need for deliberate consideration of which aspects of human existence should be preserved as AI development continues to advance. Pointing out the rapid advancements in AI benchmarks and the potential for superhuman AI in the near future, Kirchner discusses the technical challenge of aligning AI to a target and the philosophical quandary of determining what to align it to. He compares this to the IO problem in neurotech, where humans may struggle to keep pace with the increasing complexity of AI developments. While mind uploading or brain-computer interface work may help address this issue, Kirchner strongly cautions that even with direct connections to AI, humans may still face risks.

Foresight Grants

Beyond convening experts for critical discussions around the world, The Foresight Institute has created several novel funding mechanisms to directly drive further development of promising research projects to advance AI Safety.

These include dedicated programs in Neurotechnology and WBE, summarized below, as well as Security and Cryptography, and Multi-agent simulations. These programs are supported through an endowment designed to disburse annual grants, while supporting a grand prize award for achievement of WBE.

With several paths to WBE development gradually emerging, the WBE grants program is designed to encourage parallel experimentation with a multitude of approaches to support the growth of this nascent, yet heterogeneous research community.

To learn more about supporting the endowment via tax-deductible donations or getting involved, please contact Niamh Peren, Chief of Strategy at Foresight Institute: niamh@foresight.org

AI SAFETY GRANT: NEUROTECH APPROACHES

Launched in 2023, our AI Safety Grants in part fund neurotech approaches for AI Safety, including BCI and WBE approaches, as well as Lo-fi uploading approaches to produce human-aligned software intelligence. Specific areas we grant include:

- BCI approaches to improve human cognition, or compete/communicate/merge with AGIs.
- WBEs which are easier to align with human values than AGIs that share less similarities with the human brain.
- Lo-fi emulations that are more cost-effective than WBE by using extensive behavioral and/or neural data of an organism to generate models that can aid with alignment.
- Other approaches in this area.

GRANTEES WHO PRESENTED THEIR WORK AT THIS WORKSHOP

Catalin Mitelut, NYU and the University of Basel

Lo-Fi Approaches for Uploading: Developing a Turing Test for Cloning of Biological Organisms Using High-Precision Multi-Modal Behavior Modeling

Mitelut develops machine learning methods to model rodent behavior and create artificial agent models based on empirical data. His focus includes evaluating reward and value modeling, as well as analyzing adversarial attacks on behavior models.

Logan Collins, Washington University

Expansion X-Ray Microscopy

Collins proposes developing a novel method, expansion x-ray microtomography, to bring human brain connectomics within reach by achieving nanoscale resolution and vastly reducing the imaging speed bottleneck. This approach could advance understanding of AI-relevant cognitive processes like empathy, sociality, motivation, and decision-making, facilitating the design of AI more closely aligned with human interests.

Roman Bauer, University of Surrey

Computational Modeling of Neural Development

Bauer demonstrates a novel approach to WBE by employing a computational model of neural development. He aims to simulate how a mature brain can be reproduced by modeling its development from a single precursor cell.

WBE FAST GRANTS AND GRAND PRIZE PROGRAM

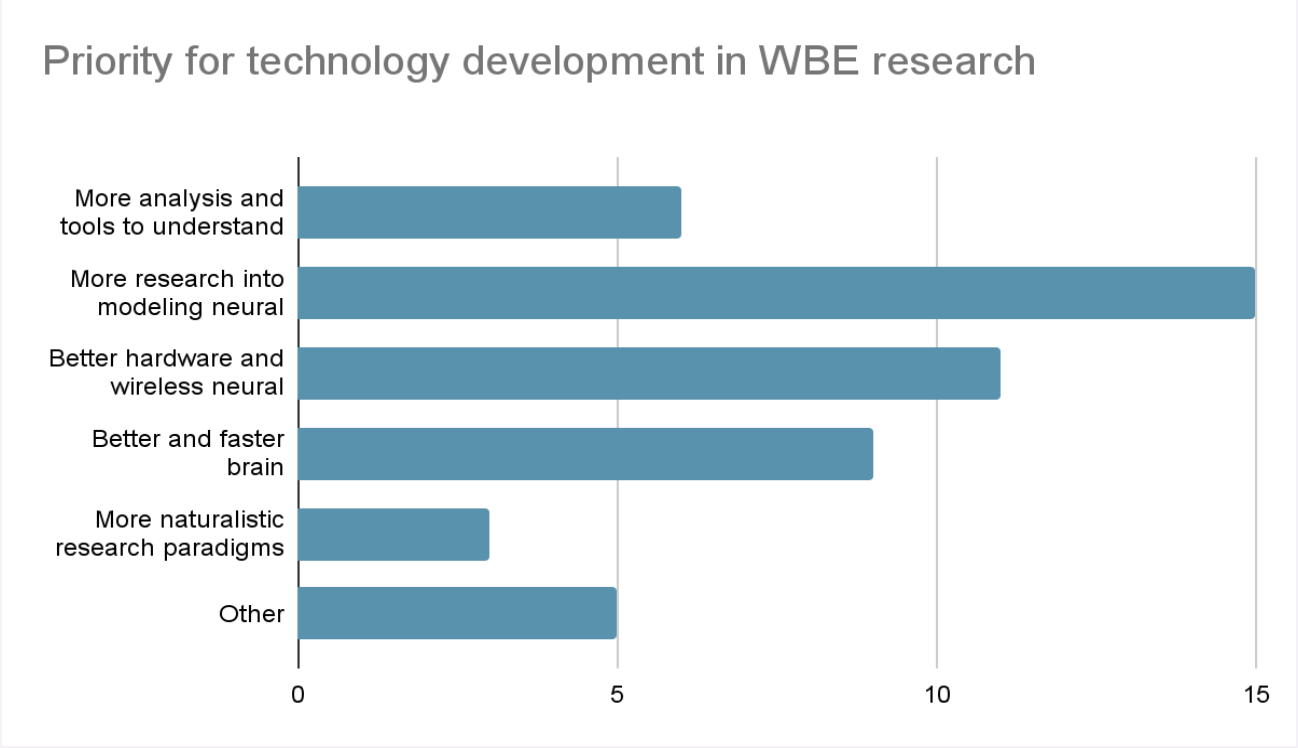
To expand upon our current field-building efforts in this space, Foresight Institute is building out a proposed $20 million WBE Fast Grants and Grand Prize program.

Given the gradual emergence of differing paths within WBE development, this program aims to encourage parallel experimentation with a multitude of approaches to support the growth of this nascent, yet still scattered research community, speeding up progress on beneficial approaches to WBE. In addition to the "traditional" approaches to WBE which focus on mapping the connectome to create a model of the brain that can be simulated moving forward, new areas include "lo-fi" approaches to emulation propose to use naturalistic behavior datasets, neural recordings, and deep learning approaches to create a good-enough model of the brain.

If you would like to support the program, either as an Advisor or Donor, our Chief of Strategy & Innovation, Niamh Peren would love to connect with you - niamh@foresight.org.

LO-FI WBE MOUSE BRAIN PRIZE

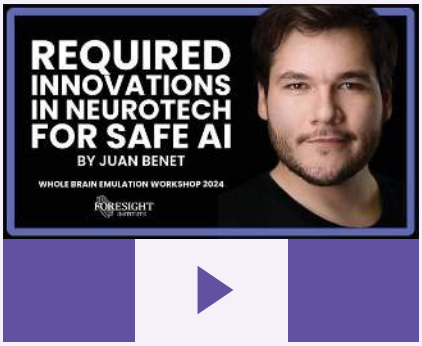
Foresight Institute is proud to announce that it will be launching a $1million dollar lo-fi WBE Mouse Brain Prize. More details about the specifics of the prize, the advisory board, and application process will be shared in coming months.



Participants weighed in on where technology developments would benefit WBE research

Comparing Approaches

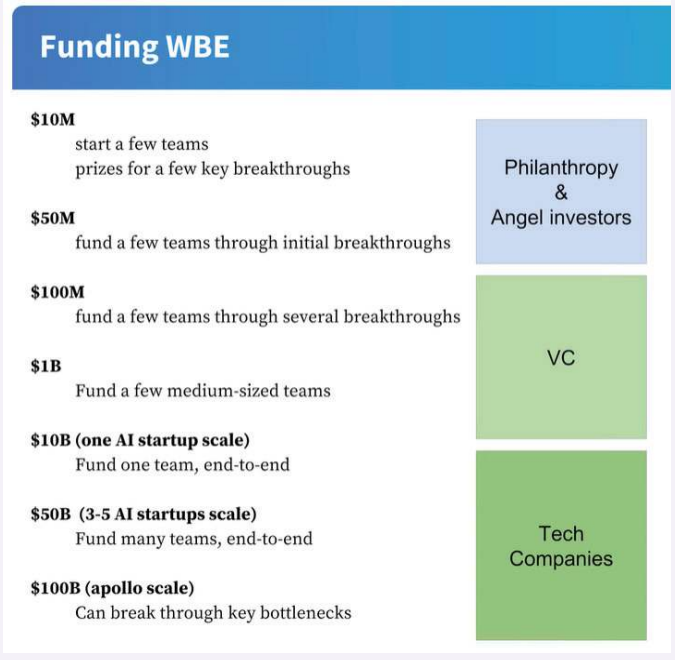
After establishing consensus knowledge, all workgroup participants united to summarize their approaches and systematically integrate various approaches, benefits, risks, timelines, budgets, and bottlenecks. Each workgroup proposed a specific project to further explore their pathway.



Required Innovations in Neurotech for Safe AI

Juan Benet, Protocol Labs

Benet emphasizes the significance of grasping computer complexity and the promise of spiking neural network models in driving AI breakthroughs. He forecasts that transformative BCIs will emerge within the next three decades, with median AI models already rivaling human performance in randomly selected tasks. Stressing the obstacles in constructing experimental frameworks for brain imaging, modeling, and simulation, he highlights the necessity for a field-wide R&D approach that enables rapid data sharing and iterative testing at scale. Drawing parallels to successful accelerated R&D programs such as the Apollo missions and Manhattan Project, Benet anticipates the coming 10-20 years to be a captivating period for AI advancements.



Comparing Neurotechnology-Based Approaches to AGI Alignment (Table1: Risk/Benefit)

Path	Benefit	Risks
Hi-fi WBE/ Connectomics	Solves alignment as humans of high enough fidelity are aligned with humans (ex: if used to design singleton WBEs that are competitive with AGI) Provides powerful way of studying prosocial neurobiological mechanisms Allows humans and posthumans to remain competitive with AIs over long term Solves other x-risks (physics catastrophes, gamma ray bursts, AGI disasters, biorisk)	Degree of technical difficulty May take longer than short AGI timelines (5-20yrs) Initial nascent approaches will lack full complexity Unknown S-Curve Even partially complete WBE might positive influence AGI dev
Lo-fi WBE/ ML and recording-driven WBE	Create human-like aligned AGI Have easily modifiable human-like intelligences compatible with current tech (backprop, transformers, etc.) Re-align hi-fi WBE to neural activity/ behavior Interpretability by correlating data to normative models Feasible also within the shorter range of AGI timelines	Lack of data Hard to fully span the space of relevant tasks/behaviors, out of distribution robustness Big behavior-neuro data + models create risks for human/AGI abuse, S-risks More variability/uncertainty for testing for S-risks (consciousness?)
Increasing human intelligence/BCIs	Smarter humans have a better shot at solving alignment. Might be necessary, if humans aren't currently smart enough. Accelerates other pathways (e.g. high speed WBE)	Smarter humans could also make unsafe AGI faster (or other dangerous tech) Concentration of power Risk of value drift Risk of mental health compromise of modified humans. (OOD brains -> unpredictable behavior)

Comparing Approaches

Path	Benefit	Risks
Neuroscience of prosociality for alignment	Solves alignment, negative alignment tax Prosocial utility may scale well/ strong uptake Could inform more informed AI better able to understand and help humans Makes AI more like us, and more likely to like us, and more capable as a result Could substantially accelerate other alignment approaches If prosocial AI does not fail safely, developing at a smaller scale first, affords potential opportunity to understand and mitigate risks	Cognitive empathy (without affective empathy) could be dangerous or not fail safely Possible lack of uptake of methods/ ideas by mainstream AI developers
WBE roadmap	Help other WBE projects, outline the safety/WBE case	Approach lock-in: Given the speculative nature of this work as well as rapid developments of enabling technology and research methods, preliminary commitments could preclude future options with greater viability.
Brain-like AGI	More interpretable and controllable AI systems. Ability to apply knowledge from neuroscience literature to aligning AGI (see Steve Byrne's work).	Faster development of unaligned AGI enabled by same research program. Exfohazard risk. Not clear how much more mech-neuroscience we need. Low certainty of success on the proposed timeframe / budget.

Comparing Approaches

Workgroup participants arrived at the following estimates for different pathways. Given the theoretical, speculative nature of these discussions, these should be considered ballpark estimates, for comparative purposes only. As such, there was not unanimous consensus on exact timelines or details. A high-level summary of the discussion across different approaches is available in Table 2.

- Hi-fi: Costs: $2-4b, 10 years
- Lo-fi: Mouse: $20m, 4 years
- BCI: $1B, > 5 years
- Prosocial: Costs: $1M, 3 years

A common refrain throughout these theoretical discussions emerged: As technology advances in areas from electron microscopy to AI-based research tools, timelines could shrink considerably, while costs could increase or decrease. Conversely, the availability of enough funding could enable simultaneous technical development and accelerate timelines.

Comparing Neurotechnology-Based Approaches to AGI Alignment (Table 2: Implementation Considerations)

	Cost	Approach	Bottlenecks	Timeline
Hi-fi	Human connectome + function + efficient running: $2B-4B One possibility to note: expansion x-ray microscopy brings cost down to ~$10M-$20M for building custom connectomics beamline	Electron microscopy, expansion microscopy, light-sheet microscopy, x-ray microscopy, NSOM face scanning Develop ANN or similar to map from structure to electrophysiology and/or Kording approach Single + multi-neuron functional activity paired with structural recording. Can incorporate all low-fi functional data	Imaging Data storage, processing, segmentation Translation problem	Construction process of first WBE: ~5 yr R&D timeline for advanced imaging capabilities, processing, and translation: ~10 years

Comparing Approaches

	Cost	Approach	Bottlenecks	Timeline
Lo-fi	Mouse $20M+ Macaque $100M Human $1-2B	Capture a variety of high throughput functional neuro and behavior data at different abstraction levels Model behavior and neuro upon central goal of safe behavior	Data acquisition Data storage Tech dev for more scalable data acquisition Modeling and compute Validation & Integration	Mouse 4 years Macaque 5-6 years Human 7 years
BCI	Germline: $10-$100M In vivo: >$100M BCI: >$1B	Genetics Germline In vivo editing BCI Connectome enhancement Exocortex	Germline: epigenomic correction (i.e. safe cloning) In vivo: delivery of nucleic acids / large molecules to brain cells BCI: high bandwidth read/write connections to brain	Germline: <5 years (to baby born) In vivo: ~5 years - 8 years (to technical feasibility) BCI: > 5 years
Prosocial	MVP: 1M for initial work merging existing homeostatic vulnerable reinforcement learning work with existing Attention Schema Theory (AST) agent work and building multi-agent scenarios that demonstrate prosociality.	Define and develop a benchmark to evaluate prosociality in multiagent AI systems Combining AST into multiagent systems toward a prosocial curriculum Evaluate models in benchmark and scale successful models	Better test environments Funding May need governance / enforcement for consistent uptake	1 year for initial implementation 2 years for full evaluation and start scaling 3 years to fully scale into larger models

PROJECT PROPOSALS

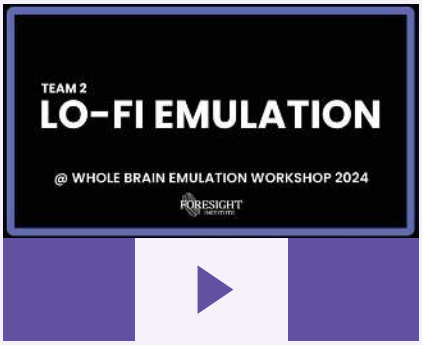
Inspired by the challenges pointed out during the keynote presentations, working groups were formed to address problems of common interests.

Hi-fi Emulation

- **Isaak Freeman**, Massachusetts Institute of Technology
- **Michael Andregg**, Fathom Radiant
- **Niccolò Zanichelli**, Università degli Studi di Parma

Working group one explores hi-fi emulations for WBE, focusing on connectomics to create realistic digital entities. They examine key areas such as connectomics, structure-to-function translation, and virtual environments. While electron microscopy currently leads, expansion microscopy with multi-beam and barcode approaches shows promise. Technical challenges include managing vast brain data and effective labeling.

Cost estimates range from $30-400 million for a mouse connectome and billions for a human one. The group discusses various approaches, including light sheet microscopy for simpler organisms and novel techniques like EMcapsulins with multiSEM. Success metrics involve developing a functional mouse connectome and its application in drug development. The timeline for mouse data is 2-5 years, with human data expected in 5-15 years. The group emphasizes pursuing multiple techniques to advance WBE research, noting that while safer than lo-fi approaches, combining methods could be beneficial.



Lo-fi Emulation

- **Bogdan Ionut Cirstea**, Independent
- **Catalin Mitelut**, NYU
- **Guillermo Vale**, Nudge
- **Jacob Cannell**, Independent
- **Johan Winnubst**, E11 Bio
- **Philip Shiu**, Eon Systems
- **Sharena Rice**, Sanmai Technologies PBC

Working group two explores lo-fi WBE and recording-driven WBE approaches for safe AI development. They propose using machine learning and behavioral data to mimic subject behavior and generalize internal structure without relying on low-level details. The group aims to bypass mechanistic models by focusing on functional representations and multimodal imaging. Key technical challenges include scaling up datasets and determining optimal resolution requirements.

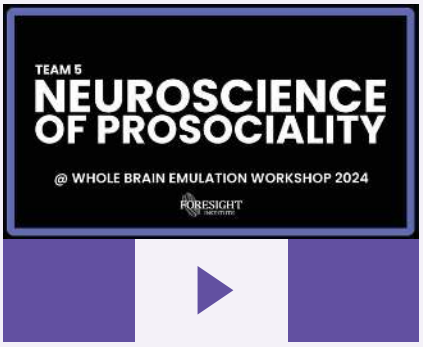
Success metrics involve developing a roadmap linking spatio-temporal scale to model quality, using behavioral metrics, and implementing causal interventions. The project timeline spans 2-4 years for non-mammalian proof-of-principle to 20 years for comprehensive human modeling, with costs estimated between $2.5-10 million for non-mammals and $5-10 million for mice studies. The group highlights potential benefits in treating neurological and psychiatric disorders, while acknowledging risks related to privacy, data misuse, and ethical implications of simulating human cognition. They emphasize the need for open data, paradigms for data generation, and neural technologies to advance this approach.

Human Intelligence Enhancement, Gene Editing and BCI

- **Benjamin Korpan**, Bootstrap Bio
- **Chase Denecke**, Bootstrap Bio
- **E.J. Chichilnisky**, Stanford Medicine
- **Tsvi Benson Tilsen**, MIRI

Working group three examines human intelligence enhancement through genetic manipulation and BCIs as potential paths to safe AI development. They explore genetic approaches, including germline editing in embryos—achievable within five years for under $100 million—and in-vivo gene therapy in adults. BCI development, estimated to cost over a billion dollars, could augment neural connectivity or expand cortical capacity.

The group identifies key technical bottlenecks, such as epigenomic correction and high-bandwidth interfaces, alongside societal challenges including ethical taboos and privacy concerns. They define success as significantly improving general problem-solving ability, potentially enhancing societal wellbeing. While acknowledging that enhanced human intelligence could accelerate unaligned AGI development, the group notes that cognitive enhancement might be crucial for addressing AI alignment challenges.



Neuroscience of Prosociality

- **Adam Safron**, Johns Hopkins University
- **Bogdan Ionut Cirstea**, Independent
- **Diogo de Lucena**, AE Studio
- **Judd Rosenblatt**, AE Studio
- **Leonardo Christov-Moore**, Institute for Advanced Consciousness Studies
- **Roman Bauer**, University of Surrey
- **Michael Skuhersky**, MIT
- **Zan Huang**, Foresight Fellow

Working group five explores bio-inspired approaches to AI safety, focusing on translating neuroscience insights into principles for an alignment curriculum. They propose combining vulnerable homeostatic reinforcement learning agents with attention schema theory to develop a prosocial curriculum. The group aims to create a multi-agent, multi-context benchmark for evaluating prosociality across various AI models. Key technical challenges include developing suitable test environments and implementing the approach in current models, including large language models.

Success metrics involve establishing a benchmark validating empathy and robustness, and creating agents with a strong aversion to harming humans across diverse scenarios. The project timeline spans three years, with an initial $1 million investment for merging existing work and building demonstrative scenarios. The group emphasizes the potential for this approach to solve alignment issues with a negative alignment tax, while acknowledging risks such as misuse of empathy by misaligned AI. They stress the importance of understanding these processes at smaller scales to mitigate risks associated with emergent consciousness and empathy in larger models.

Foresight Institute
101A Clay Street, Box 185.
San Francisco, CA 94111