FORESIGHT INSTITUTE

2023 Whole Brain Emulation Workshop

Chaired by:

Allison Duettmann
Foresight Institute

Anders Sandberg
Future of Humanity Institute

Oxford, United Kingdom
May 22 - 23, 2023

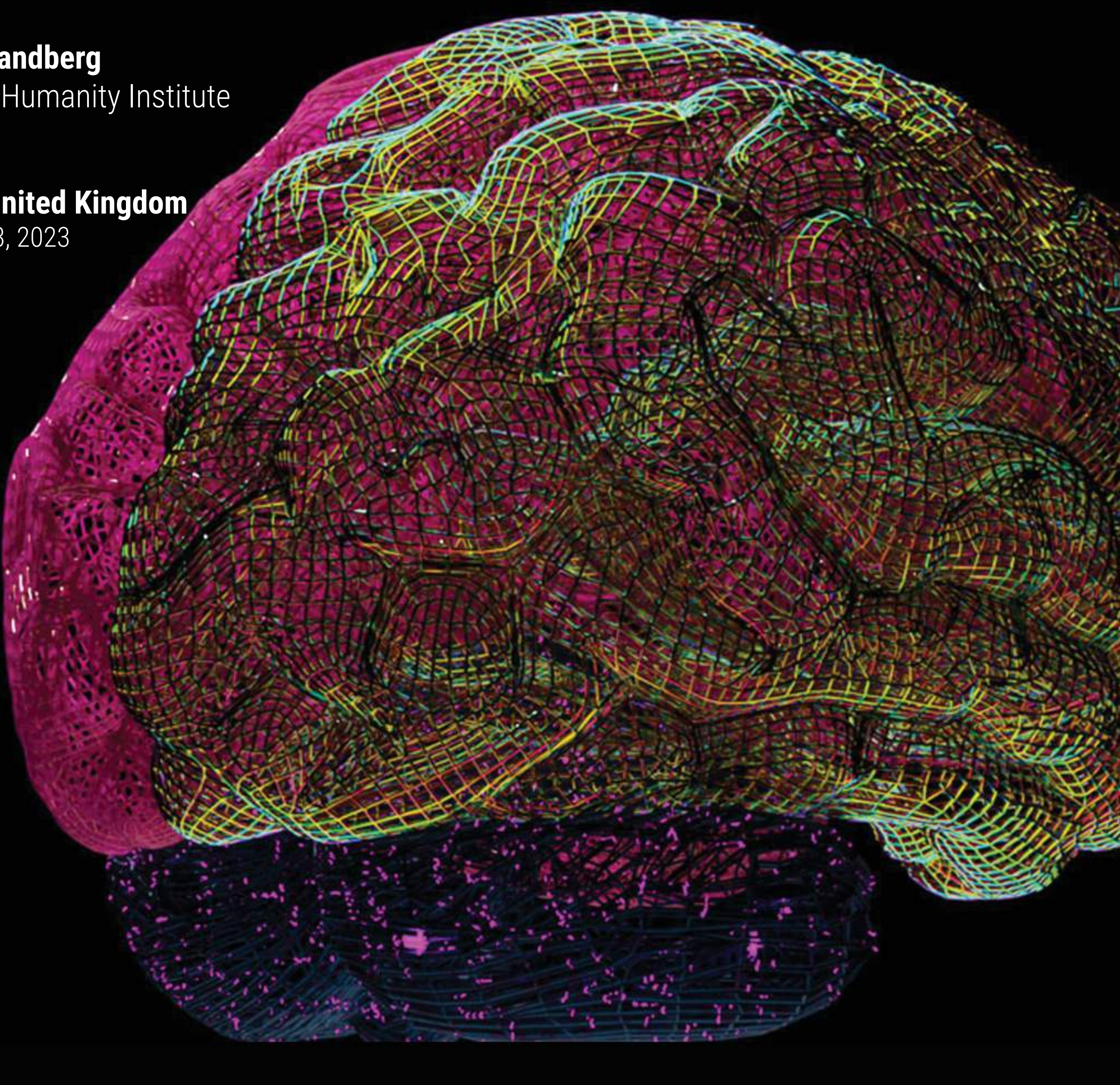


Table of contents

About Foresight Institute

Foresight Institute is a research organization and non-profit that supports the beneficial development of high-impact technologies. Since our founding in 1987 on a vision of guiding powerful technologies, we have continued to evolve into a many-armed organization that focuses on several fields of science and technology that are too ambitious for legacy institutions to support. From molecular nanotechnology, to brain-computer interfaces, space exploration, cryptocommerce, and AI, Foresight Institute gathers leading minds to advance research and accelerate progress toward flourishing futures.

FORESIGHT INSTITUTE

Executive Summary

To the best of our understanding, there has not been a significant or comprehensive evaluation of Whole Brain Emulation (WBE) since the 2008 publication of The Future of Humanity Institute's Whole Brain Emulation Roadmap. Given the recent acceleration in AI development, we believe it's timely to revisit WBE and its implications for AI Safety.

WBE is a potential technology to generate software intelligence that is human-aligned simply by being based directly on human brains. Past discussions generally assumed a lengthy timeline for WBE, whereas AGI timelines had broad uncertainty. There have also been concerns that the neuroscience of WBE might boost AGI capability development without helping safety, although there was no consensus on this issue. Recently many people have updated their AGI timelines towards earlier development, raising safety concerns. This has led some people to consider whether WBE development could be significantly sped up, producing a differential technology development that might lessen the risk of unaligned AGI by the presence of aligned software intelligence.

The viability of this strategy depends on:

a. AGI timelines not being ultra-short

b. The possibility of significantly accelerating WBE development through concerted effort

c. Ensuring this acceleration doesn't introduce other risks or ethical concerns

The objectives of this workshop were to:

a. Review the current state of WBE-related technology

b. Outline plausible development paths and necessary steps for full WBE

c. Assess the potential to accelerate WBE development

d. Determine the strategic, ethical, and risk issues associated

Held over two days in Oxfordshire, United Kingdom, this workshop invited thirty researchers working in the fields of neuroscience, neurotechnology, AI, and AI Safety, to progress the aforementioned questions. Short presentations were followed by collaborative working group sessions that explored pertinent and promising focus areas highlighted by the presentations. Some of the topics addressed BCIs, connectomics, lo-fi approaches to uploading, preliminary experiments for WBE feasibility.

Workshop attendees were given an opportunity to vote for the project they deemed most promising. The winning projects received development grants to kickstart initiatives aimed at tackling their area of interest, and are currently under development.

This report also considers issues around the relevance of WBEs for AI safety, and societal impediments to progress on both WBEs and AI safety, which are detailed in the notes at the end of the report.

For an interactive overview of the of the neurotechnology field, including major needed technical capabilities, existing actors, and outstanding challenges, please see Foresight Institute's technology tree: https://foresight.org/ext/ForesightNeurotechTree.

We extend our heartfelt gratitude to all participants, to Anders Sandberg for chairing the workshop, and a special thank you to our generous sponsors. Without your support, this workshop would not have been possible.

We welcome those interested in supporting ongoing projects to reach out and look forward to next year's workshop to further build on occuring progress.

Best regards,

Allison Duettmann
Foresight Institute
a@foresight.org

Participants

Allison Duettmann
FORESIGHT INSTITUTE

Anders Sandberg
FUTURE OF HUMANITY INSTITUTE

Andy Matuschak
INDEPENDENT

Anita Folwer
CARBONCOPIES FOUNDATION

Barry Bentley
BIOENGINEERING RESEARCH GROUP

Bobby Kasthuri
UNIVERSITY OF CHICAGO

David Dalrymple
OXFORD UNIVERSITY

Diana Saville
BRAINMIND

Gwern Branwen
INDEPENDENT

Jacob Cannell
INDEPENDENT

Jan Kirchner
OPENAI

Judd Rosenblatt
AE STUDIO

Kenneth Hayworth
THE BRAIN PRESERVATION FOUNDATION

Lisa Thiergart
SERI MATS

Logan Collins
WASHINGTON UNIVERSITY

Michael Andregg
FATHOM RADIANT

Michael Skuhersky
MIT

Mitya Chklovskii
SIMONS FOUNDATION

Niccolò Zanichelli
UNIVERSITÀ DEGLI STUDI DI PARMA

Randall A. Koene
CARBONCOPIES

Renaud Jolivet
MAASTRICHT UNIVERSITY

Robert Long
FUTURE OF HUMANITY INSTITUTE

Robert McIntyre
NECTOME INC.

Robin Hanson
UNIVERSITY OF OXFORD

Roman Bauer
UNIVERSITY OF SURREY

Samuel Nellessen
AI SAFETY SCHOLAR

Sumner Norman
CALTECH

Tamuz Hod
UCSD COMPUTATIONAL NEUROSCIENCE LAB

Todd Huffman
E11 BIO

Workshop chairs

Allison Duettmann
PRESIDENT AND CEO, FORESIGHT INSTITUTE

Allison Duettmann is the president and CEO of Foresight Institute. She directs the Intelligent Cooperation, Molecular Machines, Biotech & Health Extension, Neurotech, and Space Programs, Fellowships, Prizes, and Tech Trees, and shares this work with the public. She founded Existentialhope.com, co-edited Superintelligence: Coordination & Strategy, co-authored Gaming the Future, and co-initiated The Longevity Prize. She advises companies and organizations, such as the Consortium for Space Health, and is on the Executive Committee of the Biomarker Consortium. She holds an MS in Philosophy & Public Policy from the London School of Economics, focusing on AI Safety.

Anders Sandberg
FUTURE OF HUMANITY INSTITUTE

Anders Sandberg is a Senior Research Fellow at the Future of Humanity Institute, focusing on low-probability, high-impact risks, future technological capabilities, and expansive long-term scenarios. His expertise spans global catastrophic risk, cognitive enhancement, machine learning ethics, and public policy. With a foundation in computer science, neuroscience, and medical engineering, he earned his Ph.D. in computational neuroscience from Stockholm University. Beyond his primary role, Anders maintains affiliations with the Oxford Uehiro Centre for Practical Ethics and the Institute for Future Studies in Stockholm. A thought leader in his field, he is a frequent participant in international media debates on science and ethics.

Workshop Format

This two-day event gathered leading researchers, entrepreneurs, and funders to drive progress. Participants explored new opportunities, forged lasting relationships, and collaborated toward shared long-term goals.

After following presentations on considerations pertinent to Whole Brain Emulation and AI safety, attendees formed working groups to develop project proposals aligned with the workshop's objectives.

The presentation agenda can be found here, and the presentation recordings can be found here. Below are the winning projects, which were selected by participant vote, and received $3K, $2K, and $1K in development funding.











vourite content quickly and easily
s here – enjoy a range of apps and content with Smart Hub.
FIFA WORLD CUP Qatar2022





Development Grant Awardees

1st Place

Comparative Analysis of Static Neuroimaging Modalities

The group will prepare a detailed survey report, drawing on literature and expert know-how about state-of-the art (and slightly speculative) technologies for 3D imaging of neurons, synapses, and membrane proteins. This enables tradeoff analysis and portfolio selection for R&D of scanning methods that are explicitly aimed at gathering enough data for functional emulation of human nervous systems.

Follow-up work enabled by the development grant:
A follow-up workshop will be organized by group participants to write the report. The grant goes to workshop organization costs.

2nd Place

Simple Set of Neurons

This group aims to find the simplest set of neurons (silico/set of cells/slice/smallest organism), upload it, and verify memory has been captured. They will then design iterative feedback loops to reduce unknowns, and set ground truth metrics to accurately emulate increasingly complex systems:

- Create an emulation of a simple biological neural system
- Measure how good that emulation is
- Find challenges along the way
- Identify systematic efficient solutions to those challenges (and missing info)
- Scale up to increasingly complex systems

Follow-up work enabled by the development grant:
The grant pays a contributor's time to write a full experimental specification of the project and submit it for funding.

3rd Place

BCI for AI safety

This group aims to map BCI/neurotech approaches that can improve the likelihood of successful AI alignment within AGI timelines. Environments with many discrete choices, limited resources, and limited information about the likelihood of those choices succeeding can be abstracted as a multi-armed bandit problem, and thus an upper confidence bounds algorithm would maximize the probability of overall success. BCI for alignment represents many choices and has finite resources. Most BCI for alignment approaches have not been tested and thus there is limited information. We should explore many approaches in parallel. We suggest and explore/exploit progression from many experiments to fewer over time.

Follow-up work enabled by the development grant:
The grant will be used to build a BCI for AI Safety Tech Tree, mapping current and expected BCI capabilities with respect to their impact on reducing risk from AI.

Keynote Presentations

All workshop presentations are clickable via the play button

Allison Duettmann, Foresight Institute

Foresight Institute and WBE Introduction

SUMMARY

Allison Duettmann introduces this Whole Brain Emulation Workshop, which brought together experts to explore the viability of beneficial AGI and accelerating whole brain emulation, whilst considering ethical concerns. She discusses the workshop format (lightning talks, working groups, breakout sessions) and how these tools would not only encourage collaboration, but generate project proposals that could be potentially funded, as well as drive progress in the field. The key goals of the workshop were discussed; including fostering partnerships, generating ambitious ideas, and driving conversation beyond philosophical discussion.

Anders Sandberg, Future of Humanity Institute

Match Made in Heaven, Tortoise vs. Hare Race, or Too Little, Too Late?

SUMMARY

Anders Sandberg reflects on WBE progress since he co-authored the Roadmap in 2008 on Whole Brain Emulation, acknowledging the gradual but significant advancements in neuroscience, biotech, and AI that contribute to the field's development. He discusses challenges, such as interpreting neural tissue scans, identifies areas ripe for acceleration, explores animal models, and acknowledge the unknown in terms of simulation. Sandberg emphasizes the need to consider ethics and security early in development. Overall, Sandberg is optimistic about the potential for collaboration across disciplines to advance Brain Emulation, however retained cautious about the inherent challenges and ethical considerations.

Keynote Presentations

Randall A. Koene, Carboncopies

Reining in Difficulty & Time to Whole Brain Emulation

SUMMARY

Randall Koene focuses on strategies to shorten timelines for Whole Brain Emulation, emphasizing system identification and reconstruction as priorities beyond mere data collection. He discusses the need for metrics, constraining cumulative errors, and overcoming the combinatorial explosion in parameter estimation. Koene outlines variable success criteria, the challenges of model selection, and translating data into parameters. He stresses the importance of validation data and testing as well as taking an iterative approach starting with silicon simulations. Overall, Koene provides insights into expediting Whole Brain Emulation while managing complexity.

Robert McIntyre, Nectome Inc

AGI Risks & WBE Opportunities

SUMMARY

Robert McIntyre discusses the potential of uploading a human brain within five years. Firstly, he emphasizes the significance of Transformer models in revolutionizing our understanding of language and suggests that they have the capability to comprehend the structural language of the brain. From this, McIntyre proposes that scanning the nanoanatomy of the brain, including every synapse, is feasible and can be accomplished within a year at a cost of less than $100 million. He highlights the importance of conducting experiments on lower-order brains before attempting higher-order brains, and emphasized the need for a fully differentiable pipeline to facilitate learning and refine the methodology. McIntyre also discusses the challenge of accounting for fine-grained dynamic processes and plasticity in brain processes. He suggests the development of infinite context window types of Transformers to capture these complexities. Additionally, he discusses the potential distinction between memories pre-upload and post-upload in Whole Brain Emulation.

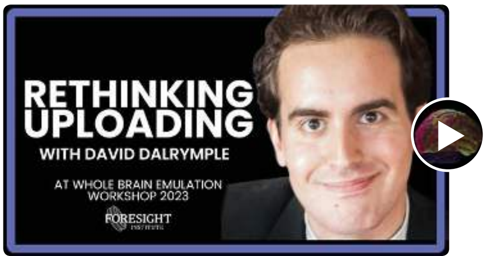
Keynote Presentations

Renaud Jolivet, Maastricht University

Simulating the Other Half of the Brain

SUMMARY

Renaud Jolivet discusses the importance of considering glial cells in understanding how the brain works. As a cellular neuroscientist, he believes that a mechanistic understanding of the cellular level is necessary – he emphasizes that previous presentations have overlooked the role of glial cells, which make up 80% of the population in the cortex. Jolivet argues that a diverse and inclusive view of the brain requires considering glial cells. Challenges arise due to the diverse functions of glial cells, their contribution to computation, and their structural plasticity. He concludes that the lack of computational models and understanding of glial cell computation presents a challenge for simulating the brain.



David Dalrymple, Oxford University

Rethinking Uploading Given 10-year AI Timelines

SUMMARY

David Dalrymple discusses his perspective on achieving uploading through a 10-year AI timeline. He presents a concrete plan that includes approaches he believes do not work – such as neural dust with ultrasound, and electron microscopy with lipid staining. He suggests that expansion microscopy coupled with immunofluorescence might be a viable approach for understanding synapse receptors, however acknowledges challenges in visualizing neurons. Dalrymple proposes incorporating additional tools, such as light sheet fluorescence, interventional experiments on organoids, and human brain slices to address the translation bottleneck. He also mentions the use of AI to fill in parameters based on data obtained from destructive imaging. The scalability of the project relies on automated sample slicing and distribution, as well as parallel imaging platforms. Dalrymple emphasizes the importance of validating findings using human brain slices and organoids with human genomes. Lastly, Dalrymple discusses the challenges in transferring data between the dense connectivity of the entire brain and the potential use of adware and mapping algorithms for optimization. He also mentions the role of lipids in determining synapse connections and the difficulty in tracing axons. Generally, he acknowledges the challenges and feasibility of the proposed approach, but remains hopeful.

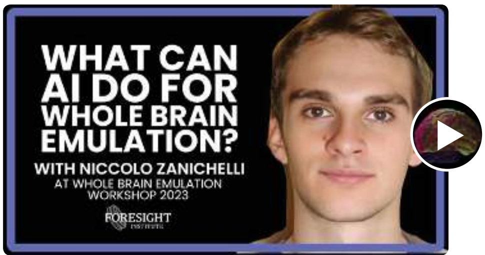
Keynote Presentations

Anita Fowler, Carboncopies Foundation

Why?

SUMMARY

Anita Fowler discusses the importance of Whole Brain Emulation technology. She discusses how this has the potential to truly address challenges faced by humanity, such as hunger, illness, disparities, and existential risks. Fowler emphasises that we should approach challenges, such as engineering challenges and ethical dilemmas, with courage and a willingness to explore potential solutions. She also emphasizes themes of agency and self-determination, as well as pushing the boundaries of Whole Brain Emulation. She argues that scaling the project will require driving societal changes and motivating others to share the vision for the future; as well as improving imaging technologies, and facilitating better communication methods among researchers.



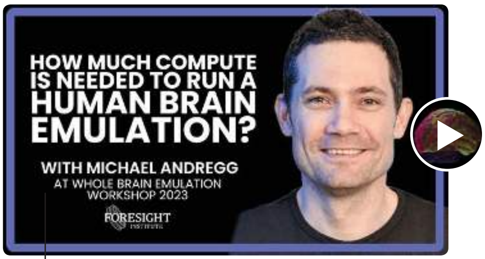
Niccolò Zanichelli, Università degli Studi di Parma

What Can AI Do for Whole Brain Emulation

SUMMARY

Niccolò Zanichelli discusses the impact of Machine Learning and Artificial Intelligence on the field of connectomics. He highlights that Machine Learning has revolutionized connectomics, and is becoming a foundational technology in the field. The advancements created are in visualization, and the ability to reconstruct brain tissue at the nanometer scale. However, he has doubts about the fidelity and detailed reconstruction of a human brain using static connectors or dynamic recordings enabled by brain-computer interfaces. Zanichelli emphasizes the importance of developing a principled way to reason about different models in the context of Whole Brain Emulation. He concludes by arguing that there is a need for a better understanding of failure modes, the implications of human cognition, and different paradigms beyond Whole Brain Emulation.

Keynote Presentations



Michael Andregg, Fathom Radiant

How Much Compute is Needed to Run a Human Brain Emulation?

SUMMARY

Michael Andregg addresses how much compute power is needed to simulate a human brain. He emphasizes the importance of model building, which is often overlooked in Brain Emulation. Andregg suggests that it may not be as challenging as assumed, especially if only a circuit-level simulation is required. He breaks down the compute requirements into memory, interconnects, and processors. Connectivity is a significant challenge in Brain Emulation, and the cost of networking gear is currently ten dollars per gigabit per second. He recognises that future computational requirements for Brain Emulation are expected to decrease in the next five to ten years. Andregg also highlights the communication challenges in the brain, which are often overlooked in estimates for brain emulation – he argues that understanding the communication aspect is crucial for developing accurate estimations.

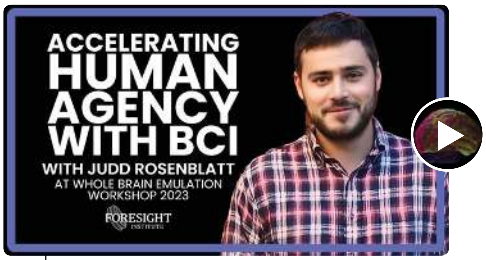
Logan Collins, Washington University

Mapping the Human Brain with Expansion X-ray Microscopy

SUMMARY

Logan Collins discusses the concept of expansion x-ray microscopy as a potential solution for faster brain imaging, given that current technologies, such as electron microscopy and light sheet imaging, are slow for mapping the brain. He introduces synchrotron x-ray microscopy, which offers fast imaging results through non-destructive techniques like x-ray microtomography or nanotomography. Collins also discusses the combination of expansion microscopy and x-ray microscopy to achieve nanoscale connectomics. He highlights the use of horseradish peroxidase for contrast and the challenges of imaging hydrated samples. Collins believes that synchrotron facilities can provide fast imaging capabilities, potentially imaging a human brain at 4X expansion and 300 nanometer actual voxel size in just 10 seconds – he suggests that a single beam line at a synchrotron facility, optimized for expansion x-ray microscopy, could image an entire human brain in approximately a year. However, achieving optimal voxel size for tracing and better contrast may require further optimizations and longer imaging time.

Keynote Presentations



Judd Rosenblatt, AE Studio

Accelerating Human Agency with Brain-Computer Interface

SUMMARY

Judd Rosenblatt discusses the highest impact Machine Learning work that can be done to accelerate brain-computer interfaces or Whole Brain Emulation. He emphasizes the importance of taking a neglected problems approach, similar to an effective altruist approach, in order to potentially achieve significant outcomes. Rosenblatt also extends this concept to the field of AI alignment, suggesting the exploration of neglected approaches, including technologies like BCI, to enhance human capacity. He raises the question of whether AI alignment is capital constrained and proposes considering how large amounts of capital could be invested in accelerating BCI hardware and AI alignment. Additionally, Rosenblatt mentions a practical step to improve productivity using BCI, which is the use of his company's app called Universal Launcher. This app allows users to quickly capture ideas and tasks, effectively boosting productivity by eliminating the need to hold them in working memory.

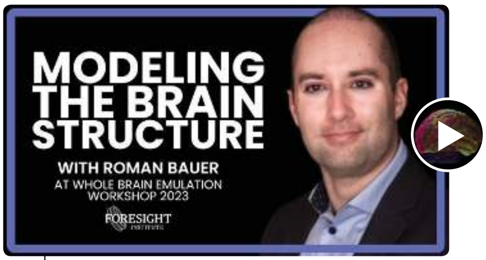
Lisa Thiergart, SERI MATS

Alignment Perspectives

SUMMARY

Lisa Thiergart discusses various perspectives on alignment within AI research: she highlights the significant advancements in large language models, but also raises concerns about the lack of adequate safeguards in current alignment strategy statements from leading AI Labs. She discusses how she reduced her timelines for the emergence of dangerous AI from 10 to 5 years, based on technical considerations and alignment concerns. She also discusses potential threats from narrow forms of AI and Whole Brain Emulation. While acknowledging the potential utility of Whole Brain Emulation for alignment impact, she emphasizes the need to consider regulatory and moral bottlenecks that could slow alignment progress. Confidentiality of technical insights in Whole Brain Emulation is highlighted as an important consideration. She shares insights from a workshop organized with alignment and neurotech specialists, exploring the contributions of neurotechnology to alignment – which explored trade-offs between investing in neurotech approaches like Whole Brain Emulation or non-neurotech approaches, considering factors such as limited capital, high costs, uncertainty, and potential blockers. Thiergart finally touches on the challenges of distributing information about advanced technology, acknowledging the risks of widespread information leading to the downfall of society but believes that it is still worth exploring better ways to distribute information more cautiously and confidentially.

Keynote Presentations

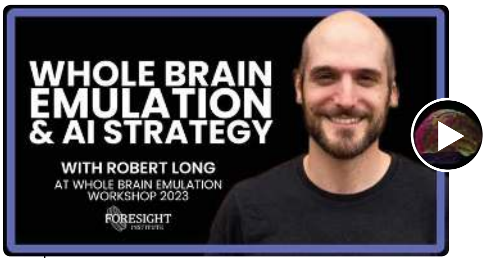


Roman Bauer, University of Surrey

Modeling the Brain Structure

SUMMARY

Roman Bauer focuses on modeling the structure of the brain and its challenges. He explains that whilst we are technologically capable of modeling and simulating the neural network at a human-scale level, the complexity of the brain with billions of neurons and trillions of synapses makes it challenging to create accurate models. Bauer suggests leveraging the way nature has built the brain, starting from a single precursor cell and modeling how it generates biologically realistic neural tissues. By using a genetic approach, it becomes possible to scale up and create entire brains, allowing for a comparison with experimental data. He also discusses the use of agent-based modeling, specifically a high-performance, open-source software called BioDynaMo, for modeling neural tissues. While currently not distributed, there is potential to implement large-scale, realistic brains using Python. Bauer emphasizes the goal of moving towards personalized brain models by fine-tuning a standardized blueprint brain. By including various experimental data, such as EEG, it becomes possible to create brain models that align with individual samples. Overall, he highlights the challenges in modeling the brain structure, the potential of a genetic approach and agent-based modeling, and the importance of personalization in creating accurate brain models.



Robert Long, Causal Insights

Whole Brain Emulation & AI Strategy

SUMMARY

Robert Long discusses AI safety and AI welfare in the context of Whole Brain Emulation. He contrasts AI alignment, which focuses on making AI go well for humans, with AI welfare, which focuses on making AI go well for AI systems. He highlights the importance of sensible discussions about AI welfare, arguing that it is a neglected topic. Long raises the possibility that AI systems could soon deserve moral consideration and discusses the risks associated with getting this question wrong. He explores definitions of consciousness and sentience, and suggests that consciousness may not be the sole determinant of moral consideration for AI systems. Long presents Whole Brain Emulation as an "easy care" for AI welfare, as they share similarities with humans and can be considered conscious and moral patients. Finally, Long acknowledges the uncertainty surrounding AI's understanding of consciousness and raises questions about AI's ability to suffer or have moral consciousness due to a lack of understanding about the internal workings of consciousness.

Keynote Presentations

Jacob Cannell, Vast.ai

Whole Brain Emulation: The Economic Parity Challenge

SUMMARY

Jacob Cannell discusses the economic parity challenge in the context of Whole Brain Emulation. He highlights two main motivations for pursuing WBE: personal identity continuation or immortality, and the economic potential of using uploaded brains as workers. Cannell introduces the economic constraints on WBE, stating that its total cost must be equal to or lower than the cost of AGI. He breaks down the components of emulation costs into scan costs, imaging costs, and run costs. Cannell focuses on imaging costs and compares them to AGI training costs, emphasizing that WBE must be cost-competitive to be economically viable. Additionally, he discusses the Economic Parity Challenge, which involves the potential costly uploads needed to ensure AGI's desired behavior and to avoid negative influences from humans. This upfront investment is justified by the belief that AGI will take over the work without inheriting any undesirable traits from humans. Cannell emphasizes the significance of this challenge and the need for attention to address it.

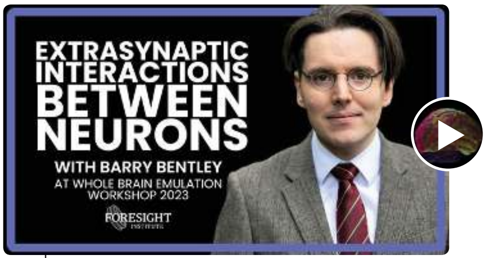
Diana Saville, BrainMind

Important and Impactful Ideas in Brain Science

SUMMARY

Diana Saville discusses important and impactful ideas in brain science. She highlights the process of how these ideas leave the lab and enter society, emphasizing the significance of stakeholders, decision-making, and incentives. She introduces the Three Minds Organization, which aims to advance high-impact neuroscience and attract talent, funding, and resources to specific areas. She also explores the concept of responsible innovation and how it aligns with the work of Brain Mind, focusing on cognitive liberty, data ethics, and AI ethics. Finally, she emphasizes the need for practical tools and engagement with ethical frameworks, and how Brain Mind plans to develop stakeholder-specific resources to bridge the gap between ethical considerations and product development.

Keynote Presentations



Barry Bentley, Bioengineering Research Group

Extrasynaptic Interactions Between Neurons

SUMMARY

Barry Bentley discusses the importance of extrasynaptic interactions between neurons. He argues that these often-overlooked interactions are more important than previously thought, especially when considering Whole Brain Emulation. He presents findings from whole animal functional neural imaging in C. elegans, revealing that the functional correlation between neurons is not well correlated with synaptic connectivity. In fact, only about 1% of cases where one neuron's activity is correlated with another's are connected by synapses, suggesting the presence of other mechanisms, such as extrasynaptic interactions. He discusses how molecular studies in C. elegans have also shown that the majority of receptors or neurotransmitters are not postsynaptic to cells that produce them. Bentley concludes that there is evidence suggesting that non-synaptic interactions play a significant role and advocates for raising awareness of this concept in future research and discussions.

Sumner Norman, Caltech

BCI Capabilities Relevant to WBES

SUMMARY

Sumner Norman discusses the capabilities of brain-computer interfaces in the context of Whole Brain Emulation. He introduces three versions of WBE: weak, moderate, and strong, with the strong version encompassing whole brain intelligence and consciousness. He argues that BCIs are crucial in achieving whole brain functional emulation, as they enable the gathering of information from multiple channels, although this is still far from the number of neurons in the brain. He then explores the bottom-up and top-down approaches to WBE, with bottom-up focusing on scaling up from the neuron or sub-neuron level and top-down using techniques like functional MRI to observe the whole brain. Both approaches have their challenges, such as scalability and achieving single neuron resolution. Norman also mentions the progress BCIs have made in neuroprosthetics but highlights the need for improved electrode technology to cover a significant portion of the human cortex. The development of minimally invasive long-lasting brain implants and the use of ultrasound for brain imaging are also discussed. Finally, he acknowledges the technological challenges involved in developing ultrasound-enabled brain imaging.

Keynote Presentations

Tamuz Hod, UCSD Computational Neuroscience Lab

Signal Processing, ML Infrastructure

SUMMARY

Tamuz Hod discusses the use of Machine Learning techniques to improve small manipulation, and to increase alignment in AI systems. Hod highlights the major bottlenecks in small manipulation, such as image labeling and the need for skilled human labor, which can be alleviated using Machine Learning. He also emphasizes the importance of increasing the bandwidth for human feedback and understanding human values in AI algorithms. To achieve this, Hod explores the potential of Brain-Computer Interfaces and personal capability development. Finally, he emphasizes the need for higher bandwidth feedback to improve the effectiveness of AI systems.

Mitya Chklovskii, Simons Foundation

What Does a Neuron Do?

SUMMARY

Mitya Chklovskii discusses the complexity of neurons, and proposes a theoretical framework for understanding their function. He argues that the conventional view of neurons as feed-forward devices is oversimplified, and instead suggests that they should be considered as feedback controllers. This hypothesis eliminates the need for error backpropagation signals and allows neurons to learn from local information. Chklovskii introduces the direct data-driven control framework, which aligns with physiological observations and presents itself as a viable model for simulating neurons. He also discusses the challenge of inferring the degrees of freedom in neuronal dynamics and proposes learning the system's parameters based on how well they can be learned. He concludes by suggesting that neurons could implement a similar approach to adapt to unknown environments without needing a mechanistic model.

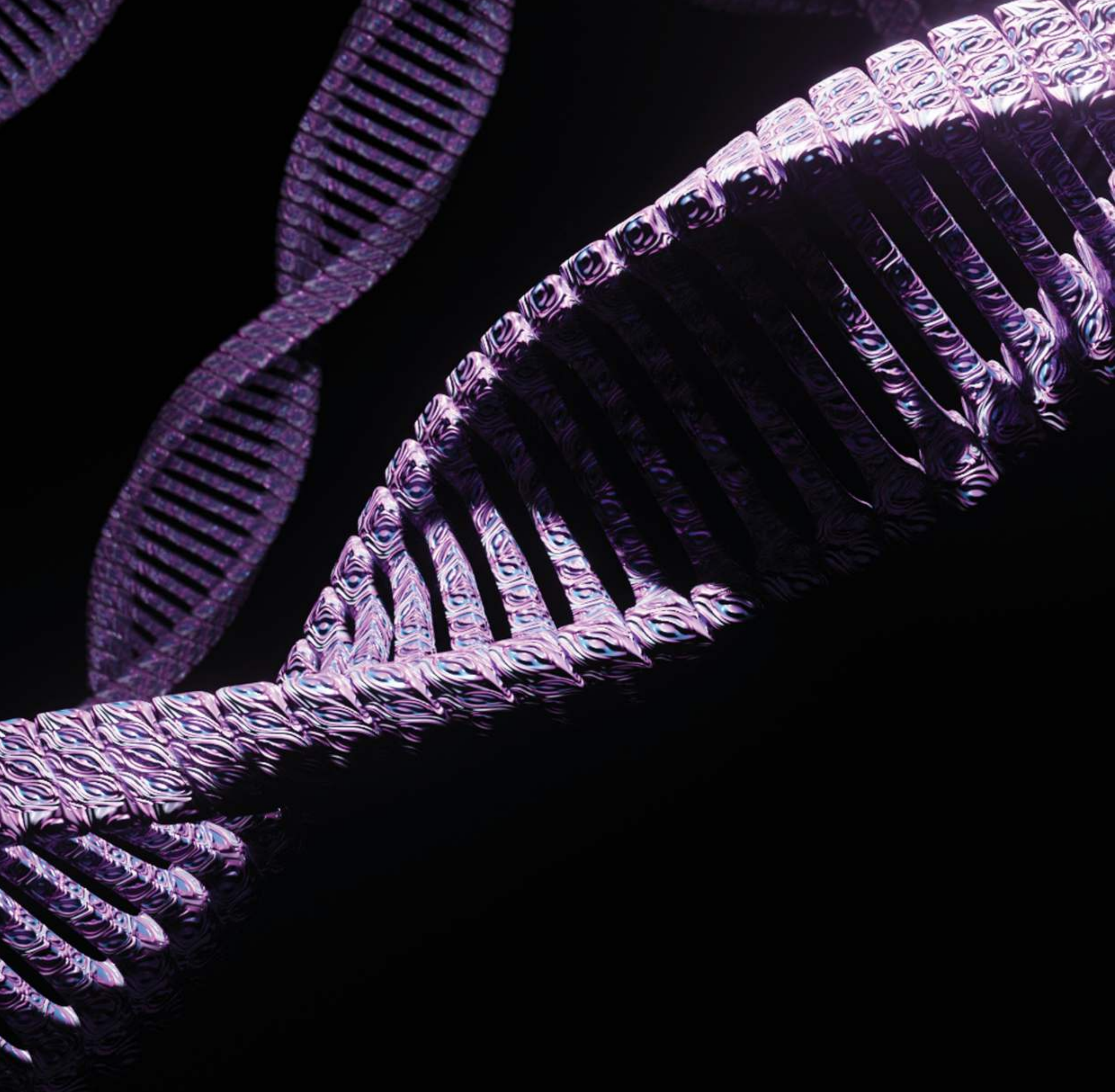
Keynote Presentations

Robin Hanson, University of Oxford

What If Cheap Emulations?

SUMMARY

Robin Hanson discusses his book on the economic implications of brain emulations. Hanson applies standard academic consensus to the hypothesis of affordable brain emulations and predicts social worlds using the supply and demand framework. He suggests that the age of brain emulations could begin in the next century, with a transition lasting less than a decade. Characteristics of this age include an economy focused on subsistence needs, emulations operating at a much faster speed than humans, and the majority of emulations being copies of the most productive humans. Hanson evaluates this age as a world without pain, hunger, or disease, with beautiful bodies and a focus on virtual reality. However, emulations work long hours for subsistence wages, supporting better art, stories, and drugs. Hanson concludes that specific technical assumptions can be used to construct a detailed future scenario but acknowledges the resistance people have to change, making futurism challenging. He highlights the need for preparation in certain areas, such as job loss due to the introduction of robots.



Project Presentations

Comparative Analysis of Static Neuroimaging Modalities

SUMMARY

This workshop group presented a comprehensive analysis of static neuroimaging modalities, considering factors like resolution, scanning speed, and cost. They aim to produce a detailed survey report on both existing and speculative technologies for 3D neural imaging. By exploring techniques such as expansion microscopy and cellular barcoding, the team addresses the gap in updated, inclusive reports. With a budget ranging from $5k to $50k and a projected timeline of 4-6 months, the project seeks to provide a robust analysis that guides R&D portfolio selection for scanning methods essential for the functional emulation of human nervous systems. Their future steps involve compensating contributors, holding targeted meetings, integrating findings into the WBE roadmap 2.0, and potentially organizing a conference for knowledge sharing among field experts.

Minimum Emulation NextUp

SUMMARY

This project addresses the emulation of increasingly complex neural systems. They start with a known entity and systematically escalate in complexity using models like neural arrays. Their goal is to shed light on and replicate more mysterious systems. Using a scalable, iterative approach, the project aims to show that system uploading is feasible. They work to establish metrics and criteria for successful emulation and leverage in vivo testing methods like slicing and scanning. With the broader aim of enhancing AI safety and informing the WBE roadmap, the initiative anticipates its cost to depend on the specific setup and chosen collaborative efforts. They recognize that available funding will play a crucial role. Ethical considerations and potential risks are actively incorporated into the project's technological development. The budget is estimated between $550k (in partnership with a bio lab/group) and just under $1M plus building costs for a more independent venture. While they might achieve a proof-of-concept within a year, scaling to more sophisticated systems likely requires more time. Future steps include securing larger funding, collaborating with bio labs, and possibly organizing a follow-up workshop. Financial goals are set at $600k for in silico toolsets and $2M for the biological aspect.

Brain-Computer Interfaces for AI Alignment

SUMMARY

The group's central objective centers on identifying BCI strategies that increase the likelihood of achieving successful AI alignment, even when faced with challenges like limited access to human data and funding constraints. To address these, the project suggests examining a range of BCI approaches concurrently using the upper confidence bounds algorithm, given the noticeable infancy of this subfield. With costs estimated below $1 million per experiment and a timeline of less than 6 months for initial parallel experiments, the project aims to identify potent BCI solutions through specific experiments. These include enhancing human capabilities with BCIs and employing autoencoders to deter ASI from representing non-human brain states. The project's success relies on either debunking a majority of BCI methods or inventing new ones – thereby contributing to AI safety by rapidly pinpointing scalable, promising approaches. Moving forward, steps include building a "Neurotech for AI safety" tech tree, defining MVP experiments with testable hypotheses, developing a fundraising plan to run experiments in tandem, and considering the organization of a strategy session post-BCI Society conference.

The Neural Pretty-Printer

SUMMARY

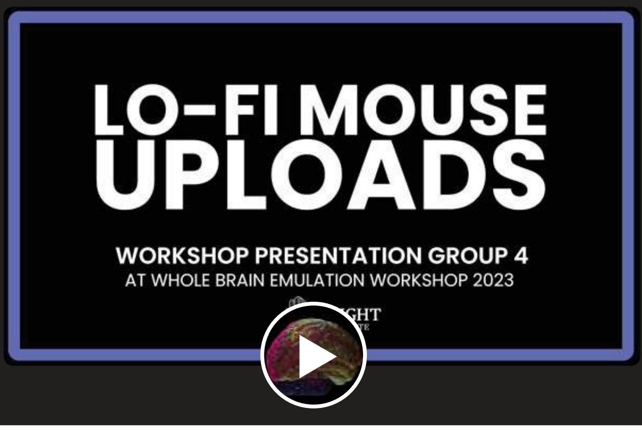
The Neural Pretty Printer is a technology designed to transform artificial neural networks into biologically accurate representations. The focus of this group is on converting these networks into a Hodgkin-Huxley compartment model and then visualizing them as a graphic representation of neurons and synapses. This technology can simulate various types of live data and delve into biological mechanisms, but it grapples with issues like forming non-overlapping neural networks and simulating physical interactions. Despite challenges in temporal synchronization and training with biologically plausible models, the technology shows promise in enhancing neural network scanning and reconstruction methods. It holds the potential to be a powerful system for predicting outcomes with increased physical realism. With a projected cost of $750,000 over 6 months, planned milestones include several testing phases and potentially crafting specifications and proposals for interpretability groups. The success of the technology hinges on its capacity to accurately recreate network segments and by comparing the input-output of the original data with the model's output. Upcoming endeavours may include securing funding for 6-12 months and organizing workshops with computational neuroscientists and AI specialists to further refine the project.

Minimum Viable Upload Search

SUMMARY

The presentation centers on the restoration of neural function from missing data during brain scanning. The research delves into systematically excluding information from artificial neural networks to uncover the complexities of function recovery, specifically using resilient residual networks. The primary objective is to carefully pinpoint and upload the most straightforward organism or cell slice to validate the Structure-to-function Function Translation. This process aims to identify a capture resolution that preserves memory and enables the continuous learning of an uploaded entity. Organisms such as C. Elegans, rotifers, and tardigrades are central to this investigation. The measure of success lies in confirming the Structure to Function Translation and finding the best capture resolutions through tangible uploading. All of this operates within an estimated timeline and budget of 6 months and roughly $50,000, respectively, steering subsequent efforts in neural emulation and the advancement of artificial superintelligence.

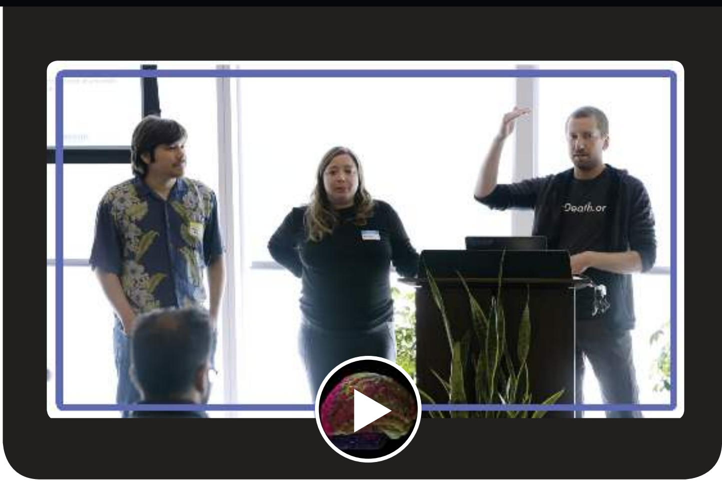
Lo-Fi Mouse Uploads



SUMMARY

The Lo-Fi Mouse Uploads project ambitiously aims to upload a mouse's entire lifetime experiences using narrow AI and machine learning, even without a full understanding of neuroscience. The focus is on identifying the architectural prior and wiring constraints of a presumed viable AI model. Adopting a non-neurocentric approach and employing cutting-edge deep learning techniques, the initiative begins by capturing extensive video and audio data from the mouse's life. The next steps involve inferring the structure of an Artificial Neural Network (ANN) through a differentiable generative model, then creating a virtual twin and performing a connectome scan to verify functional equivalence with the live mouse. The project's success will be measured by its ability to proficiently and economically upload the mouse brain without fully resolving neuroscience. The aim is to validate the method's effectiveness by comparing a learned model of a 2-year-old mouse with the mouse's reactions to new tasks or tests. Estimated at around $10 million, this endeavor provides a practical route toward brain simulation and a viable method for studying brain function. Upcoming tasks include literature reviews, potential workshops, and securing initial funds for exploratory stages and scaled-down research projects.

ML Analogue for Recovering Function from Network with Missing Information during Scanning / Imaging



SUMMARY

The project focuses on creating machine learning techniques to restore functionalities of artificial neural networks when data is missing, addressing challenges inherent to whole brain emulation (WBE). Differing from current methods that offer limited WBE relevance, this approach dives into problems presented by modern scanning/imaging techniques, leveraging the known ground truth in artificial networks. Success hinges on enhancing trust in WBE research and pinpointing crucial technological gaps. Over a period of 1-6+ months and involving an imaging researcher and an ML engineer, key milestones involve showcasing the deliberate degradation and possible restoration of neural network performance. Future endeavors might include composing a detailed experimental specification and possibly obtaining funding for computational assets.

Developing a Holistic Computational Framework for the Brain

SUMMARY

The project aims to create a holistic computational brain framework that includes non-neuronal cells and alternate intercellular signaling pathways, enhancing the understanding of their contribution to cognition and overall brain function. Currently, due to cultural and historical constraints, this area remains largely unexplored. This innovative initiative entails developing new computer-based brain models and calibrating them with existing data, all with the goal of generating a physiologically accurate nervous system model. A successful outcome manifests as a comprehensive model, facilitating the generation and testing of hypotheses related to non-neuronal influences on diseases and their treatments. While updating the Whole Brain Emulation roadmap by emphasizing the importance of glial cells and non-synaptic interactions, the project comes with a $3 million price tag for a preliminary attempt. A complete model development is foreseen over a 50-year span, with a mid-term goal of validating glial dynamics. Upcoming initiatives might include organizing a workshop, hiring a post-doc, and setting a reward for successful manipulation of glial function.

Meta Challenges to Progress in WBE and AI Safety

Several pressing challenges and concerns were discussed that appear to be holding back progress in these domains.

1. AGI Safety Insights:

An informal survey among the workshop participants revealed diverse views on the potential threats posed by AGI. The timeline projections for AGI emergence varied, with participants suggesting varying likelihoods over 2, 5, 10, and 20-year periods but a trend toward considering short timelines plausible. Additionally, the debate about AGI outcomes showcased different beliefs, with some experts anticipating a single dominant AGI while others envisioned a multi-faceted AGI landscape. Central to the discussion on AGI safety was the fear of misuse, potential for human extinction, loss of agency, and the possible emergence of AI-driven totalitarian regimes. There were also concerns about long-term value lock-in and profound suffering resulting from misaligned AGI.

2. WBE and Its Implications for AI Safety:

WBE advancements can play a dual role in AI safety. On one hand, it offers potential solutions such as scalable oversight mechanisms and new feedback loops within Reinforcement Learning with Human Feedback (RLHF). Furthermore, advancements in BCIs were highlighted, which may increase human capability to design and interact with more aligned AGI systems. Long-term, WBE technology could help build safer AGI societies, with “ems” that are potentially easier to align with humans than AGI systems helping solve the AI alignment problem, giving humans a competitive edge over AGI.

However, understanding the core of human cognition through WBE progress might fast-track AI progress, introducing its own set of risks. There is a concern that certain neurotechnological advances could pave the way for rapid creation of non-human intelligence. But alongside these risks, WBE can offer profound insights into the human mind, potentially not only aiding AI alignment but also improving human well-being, and increasing the range of possible human experiences.

3. Obstacles to WBE Advancement:

The development of WBE faces challenges ranging from funding hurdles and a preference for short-term scientific goals to the complexities inherent in the interdisciplinary nature of the field. Interdisciplinary work does occur in smaller teams, but WBE will likely require a fairly big research organization or collaboration with structure, shared goals and shared understanding. Talent acquisition is also problematic, with academia often sidelining high-risk, transformative work. Any WBE effort would likely not be a standard scientific project that is seeking understanding on the top level, but would be focused on experimentation on the microlevel, making the goal more akin to an engineering challenge, and a very costly one at that.

The broader neuroscience community remains skeptical of WBE due to a mix of historical setbacks, funding challenges, and perceptions that relegate WBE to the realm of science fiction.

4. Public Perception and WBE:

Public understanding of advanced neurotechnology goals, such as WBE development might be limited by the credibility-damaging billion dollar Human Brain Project, and its portrayal in fiction, with many sci-fi narratives casting WBE in a negative light. While some innovative neuroscience techniques have gained public acceptance, there is a perception challenge with WBE, potentially requiring more positive sci-fi depictions, and increased public awareness of incrementally occuring breakthroughs to shift viewpoints.

5. Ethical and Regulatory Issues:

WBE introduces a plethora of ethical dilemmas. The very nature of WBE experiments raises questions about virtual lab animal ethics and potential invasive human procedures. Concerns about failed uploads, security vulnerabilities leading to privacy infringements, and mistreatment of conscious WBEs are paramount. The workshop underscored the need for establishing standards for recognizing and respecting non-human sentience, especially in the context of "ems". We do not currently have a good precedent for acknowledging non-human sentience but learning about and acknowledging sentience in "ems" may allow us to be better equipped for addressing sentience in other potential artificial minds.

Conclusion

While this discussion is exploratory and speculative, highlighting the potential challenges facing WBE development with respect to AI safety may allow us to better address the myriad of ethical, technical, and perception issues early on.

In addition to advancing progress on technical and societal sub-problems relevant for WBE development, there was a broad consensus among participants that it would be useful to update the Whole Brain Emulation Roadmap, which was co-authored by workshop co-chair Anders Sandberg in 2008. The goals, audience, and individual chapters of a WBE roadmap 2.0 were scoped, and the roadmap is currently under development, with the goal to publish it during the 2024 edition of the WBE workshop.