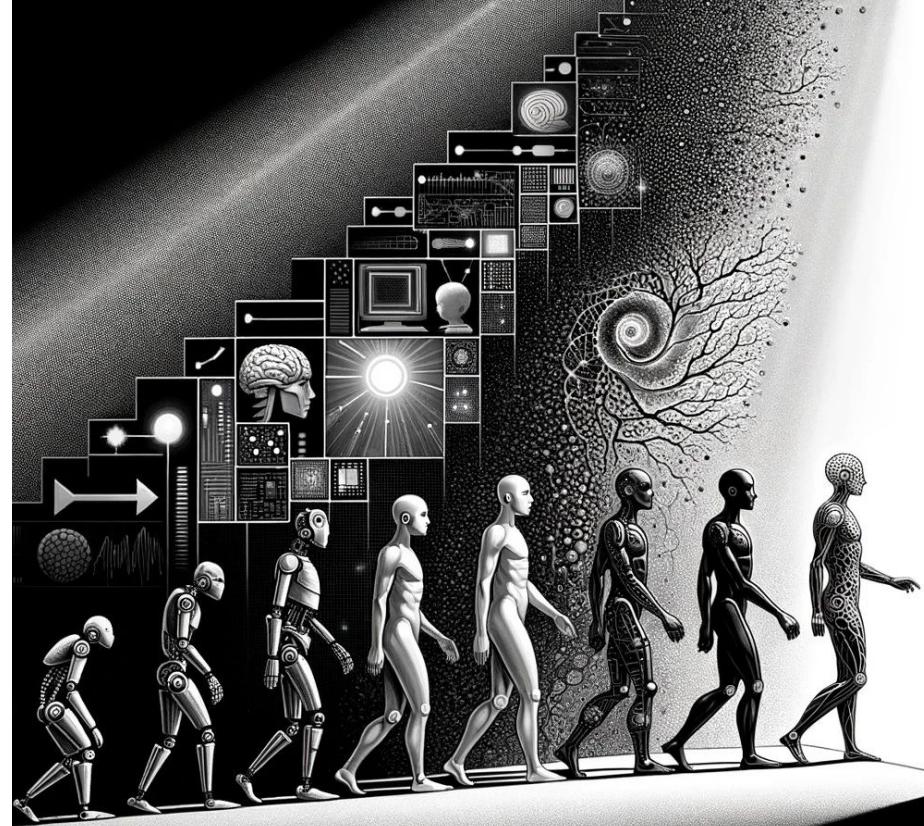


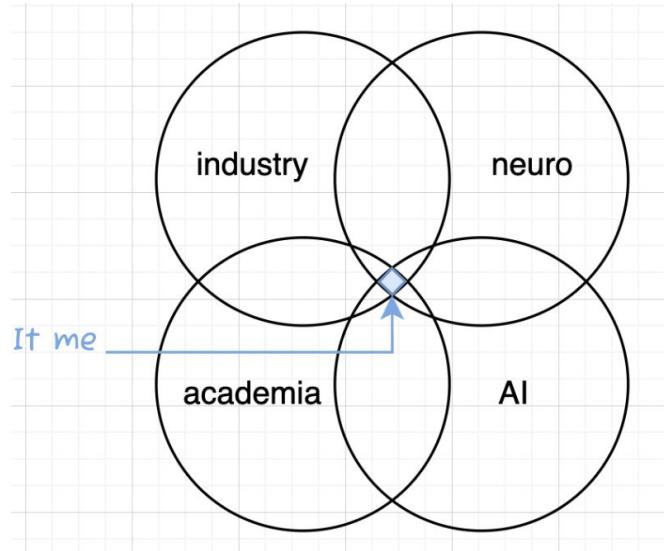
NeuroAI for AI safety

Patrick Mineault, PhD
Amaranth Foundation



About me

- PhD in comp neuro vision at McGill
- Data scientist at Google
- Brain-computer interface engineer at Meta
- Senior ML scientist at Mila
- NeuroAI researcher at the Amaranth Foundation focused on NeuroAI safety





Funding ambitious research in
longevity and neuroscience



- P Can you summarize the arguments made in this blog post I wrote about opportunities and pitfalls for foundation models in neuroscience that I will give to the NIH neuroethics working group? Use a bullet point format

Here's a summary of the key arguments made in your blog post about opportunities and pitfalls for foundation models in neuroscience:

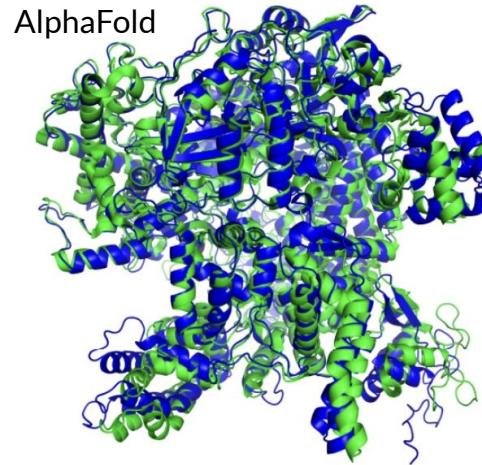
Opportunities:

- Foundation models are composable, enabling more rapid progress than training from scratch for each task
- They can be used for multiple downstream tasks via simple fine-tuning, moving from bespoke models to a "library model" approach
- While training from scratch is resource-intensive, downstream uses require less compute, making them more accessible
- They can leverage unlabeled data, which is far more plentiful than labeled data
- Foundation models have been proposed for data formats that have resisted conventional large-scale machine learning (e.g., graph data, spikes, transcriptomics)
- They represent a shift to a data-centric view of AI, emphasizing the importance of data over model architecture
- Potential to accelerate both discovery in neuroscience and applications like brain-computer interfaces

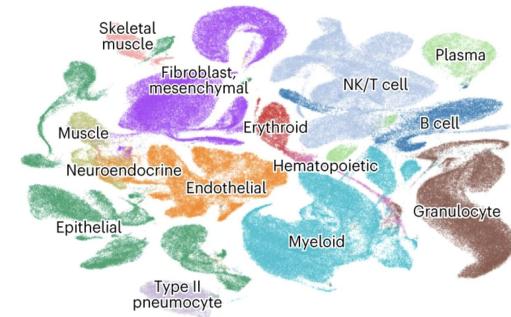
Pitfalls:

- Risk of amplifying societal biases, depleting autonomy and agency, and raising privacy concerns when used in sensitive human health-related applications
- Inscrutability of large-scale trained models, making it challenging to fully characterize biases and identify instances of shortcut learning
- The paradox of automation: as AI systems become more integrated into decision-

AlphaFold



Rosen et al. 2024





Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Geoffrey Hinton

Emeritus Professor of Computer Science

Yoshua Bengio

Professor of Computer Science, U.蒙特利尔

Demis Hassabis

CEO, Google DeepMind

Sam Altman

CEO, OpenAI

David Krueger

Assistant Professor of Computer Science

Jacob Steinhardt

Assistant Professor of Computer Science

Jaan Tallinn

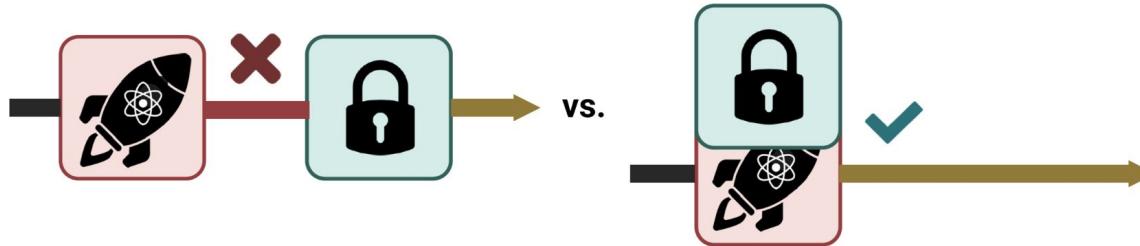
Co-Founder of Skype

Vitalik Buterin

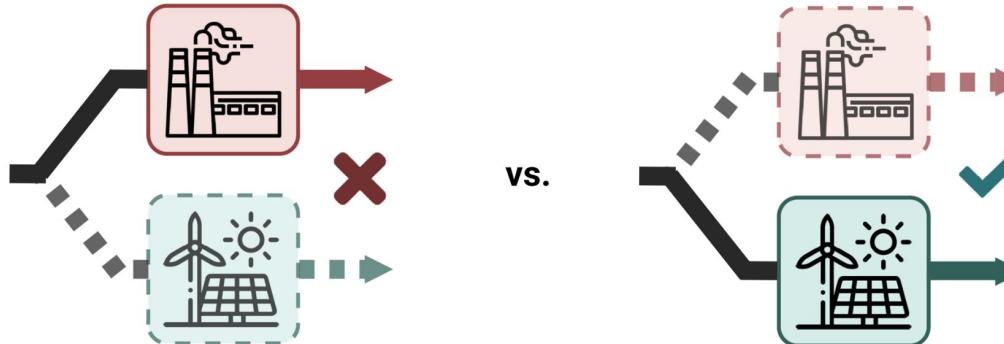
Founder and Chief Scientist, Ethereum

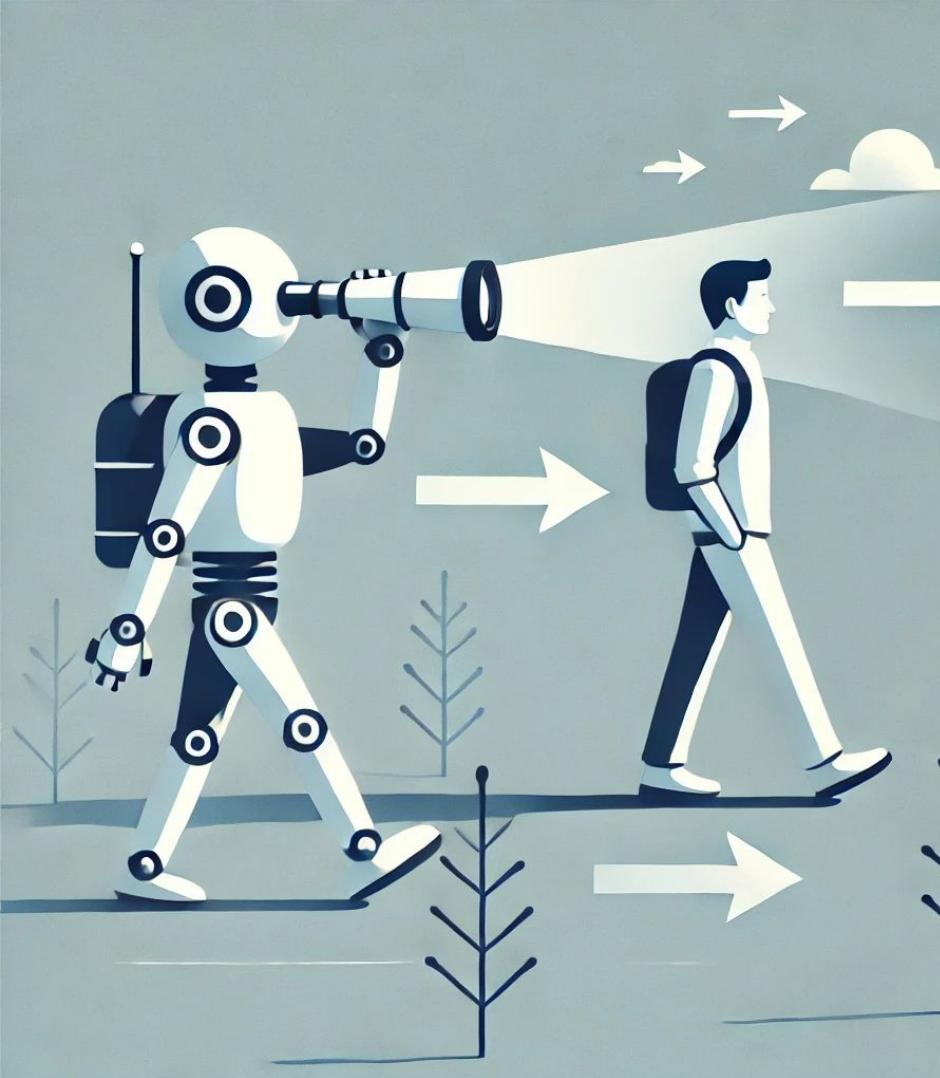
Mitigating AI risk

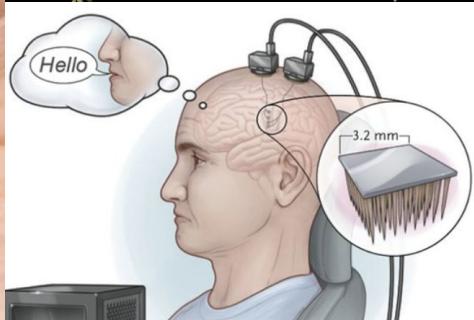
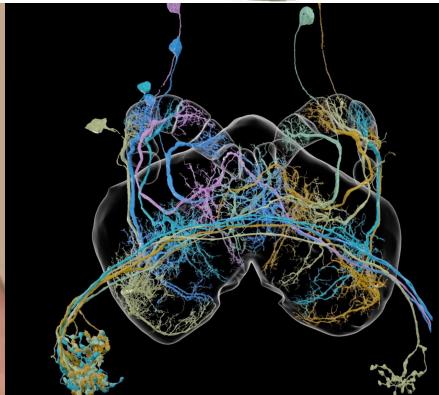
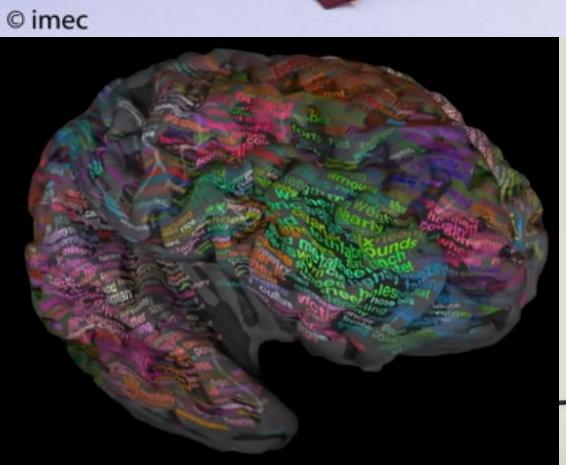
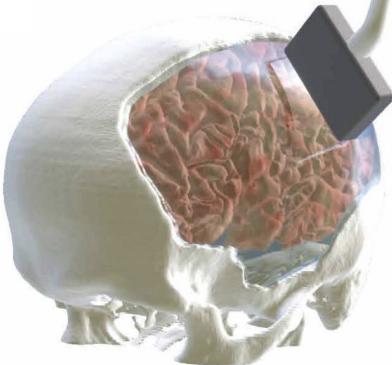
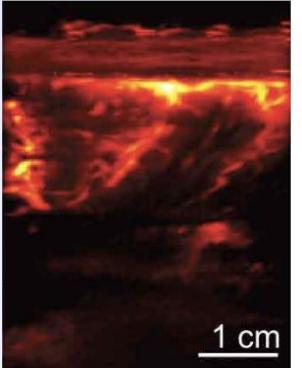
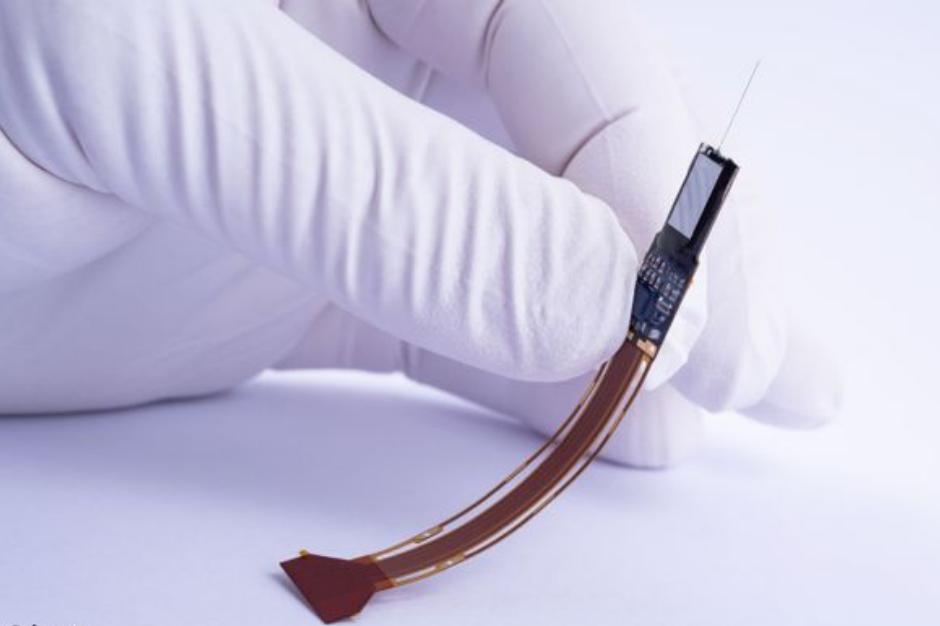
Safety technologies sooner relative to risk-increasing technologies



Substitute technologies instead of risk-increasing technologies



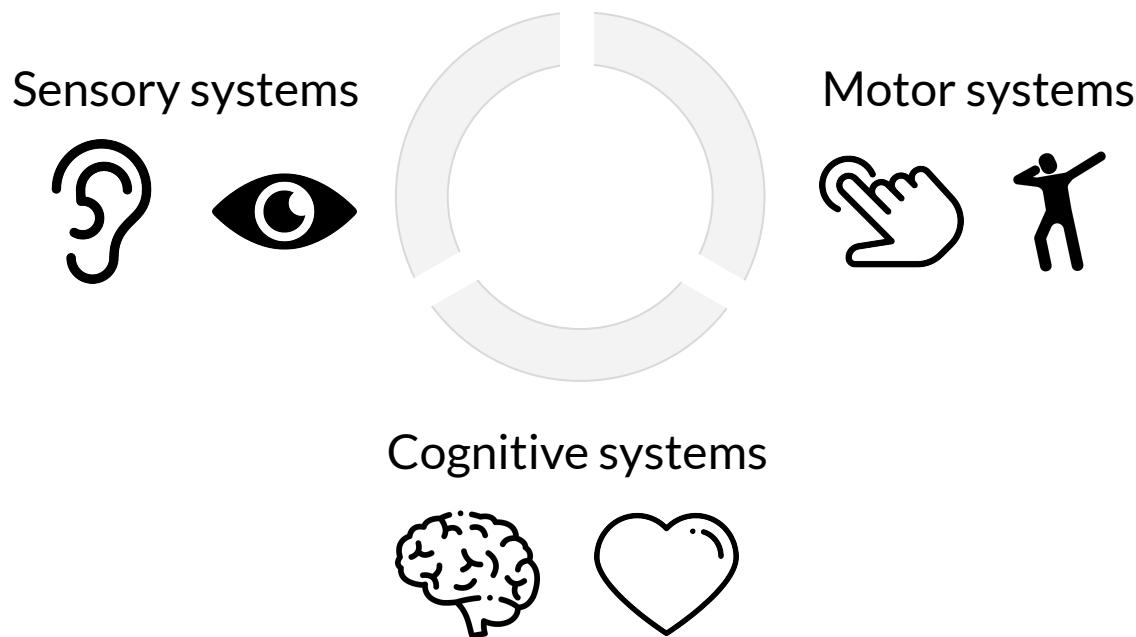


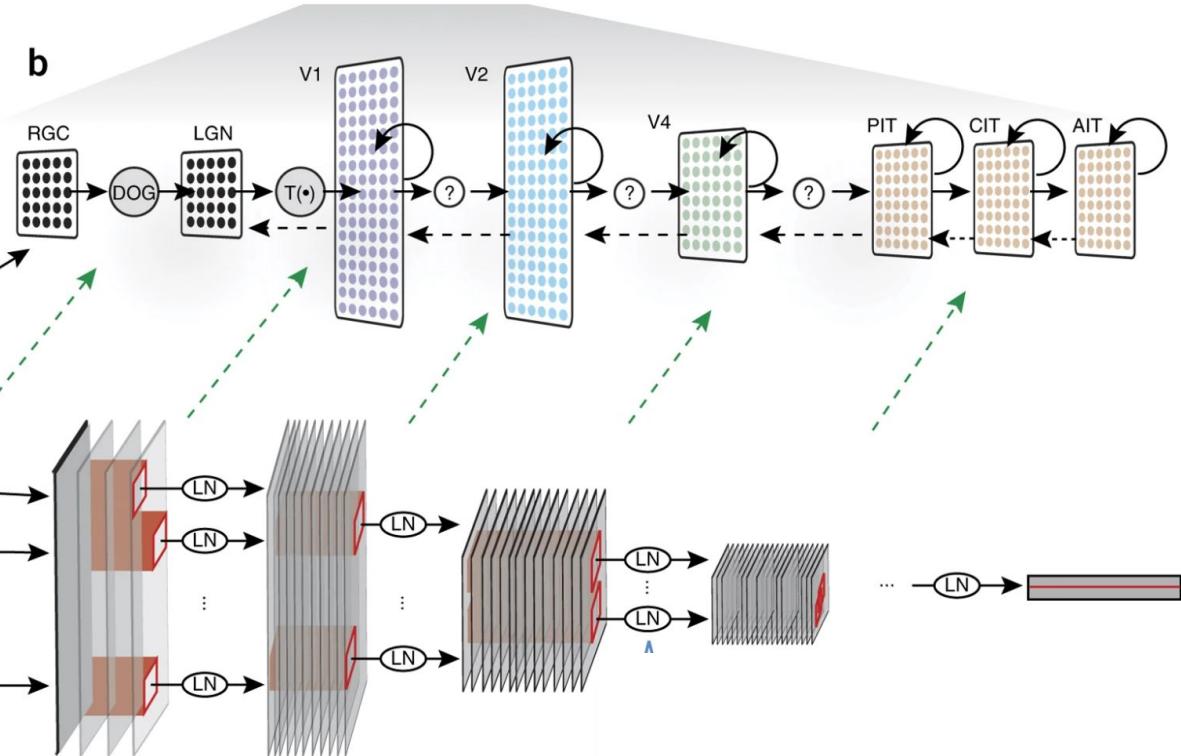
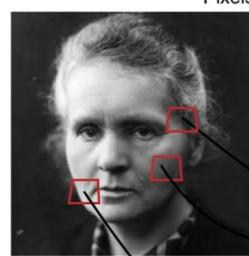
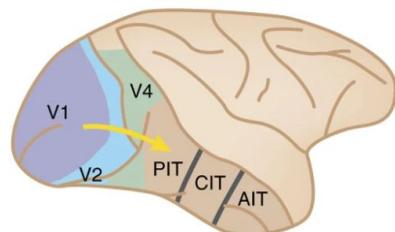
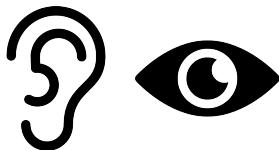


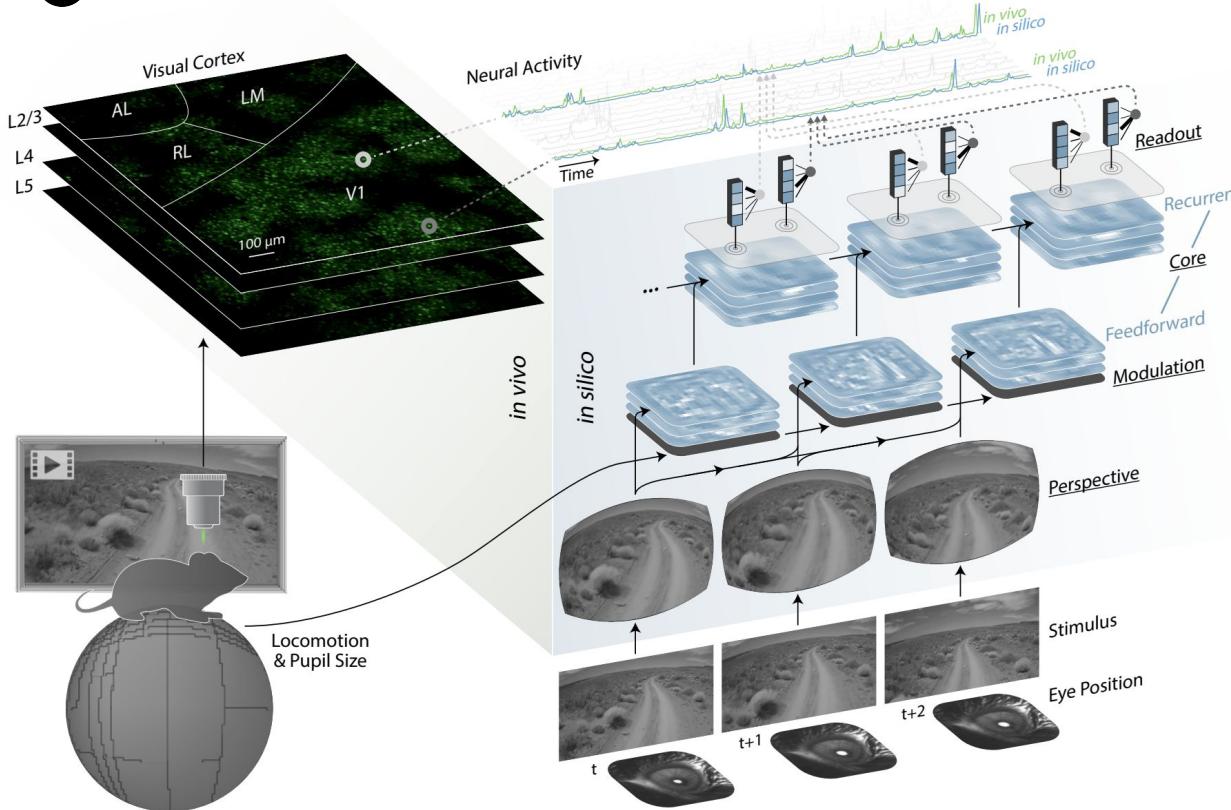
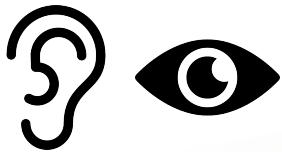
Why would brain imitation AI be safer?

- Humans know how to explore safely
- Humans are robust (adversarial, out-of-distribution)
- Humans cooperate*
- Humans know how to reason about humans
- Our world is built for humans
- Those capacities are a function of our brains
- Brain data represents a richer set of constraints than behavior

A path toward brain imitation learning

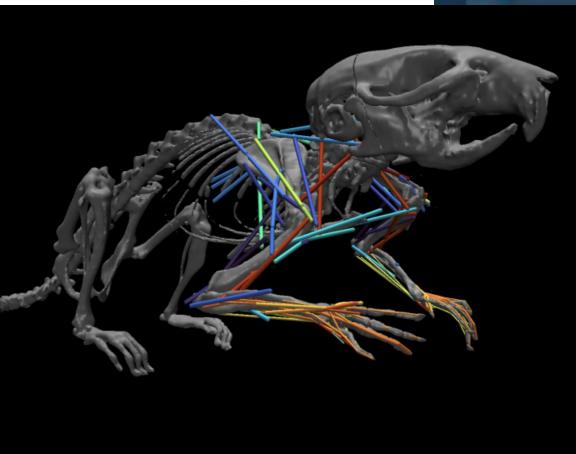
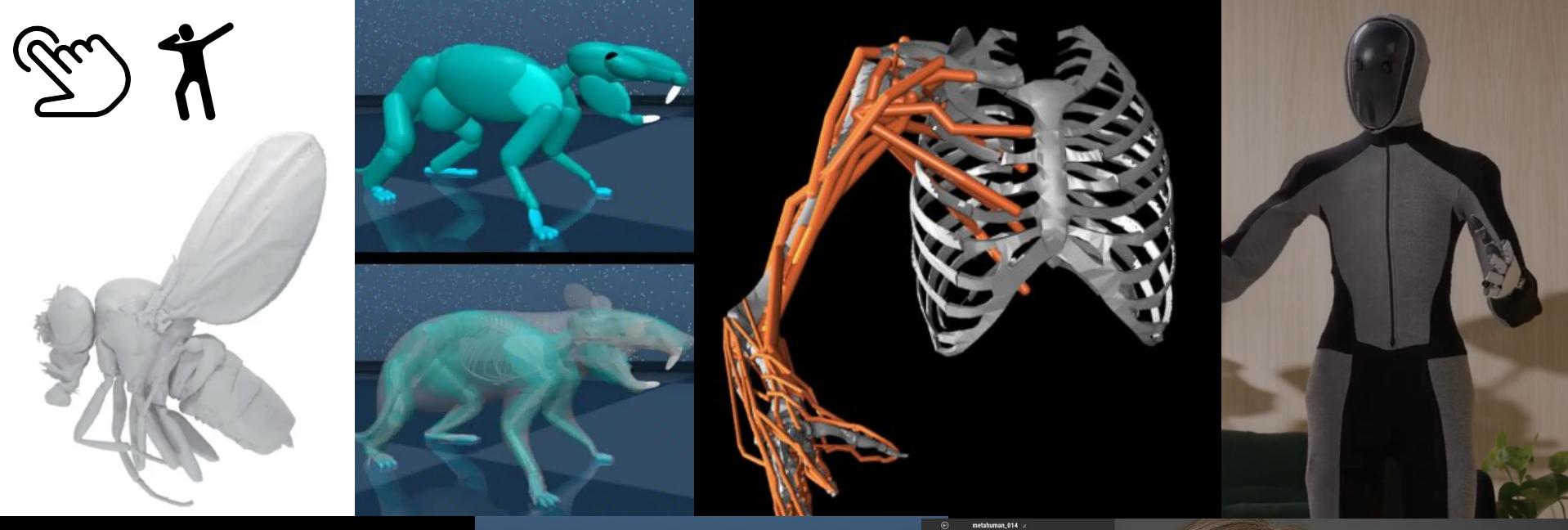






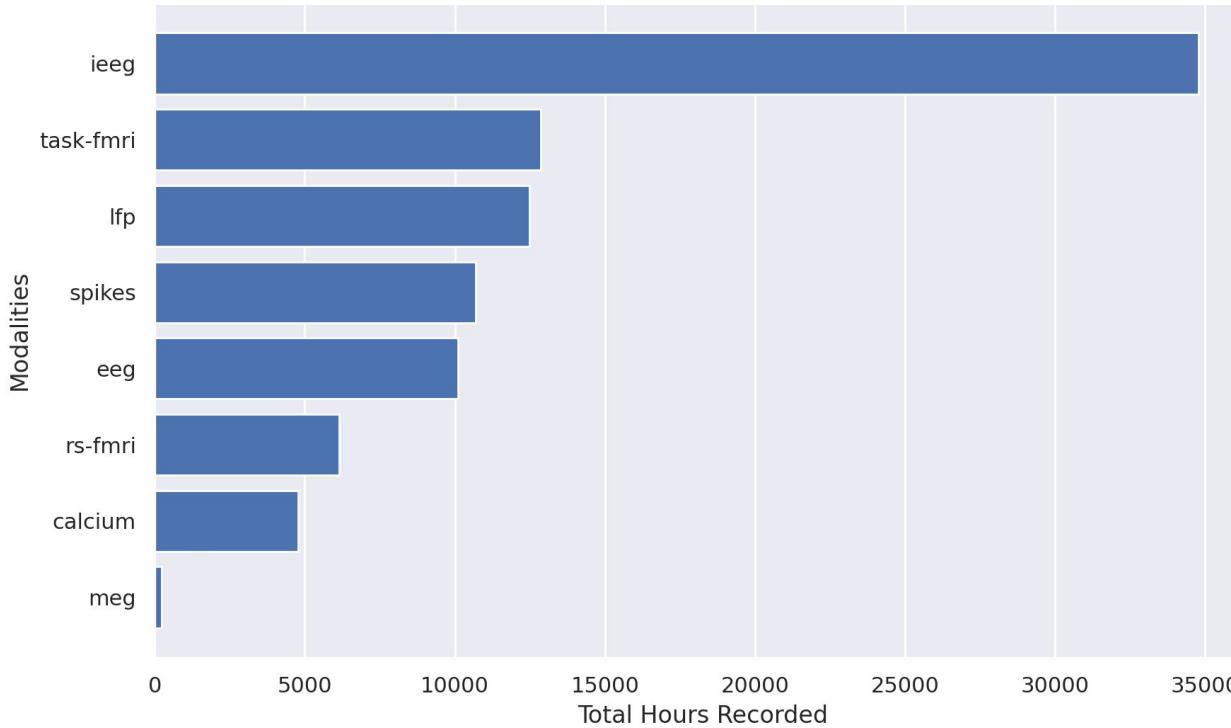
The Enigma Project

A quest to decrypt
the neural code





Total hours freely available across modalities



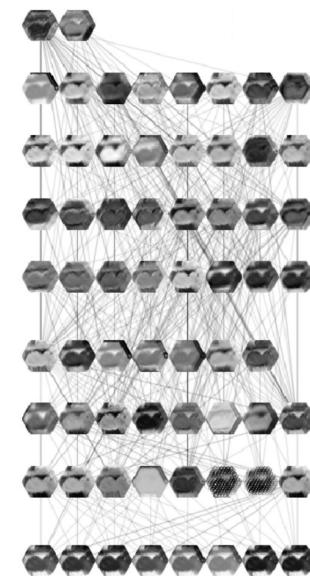
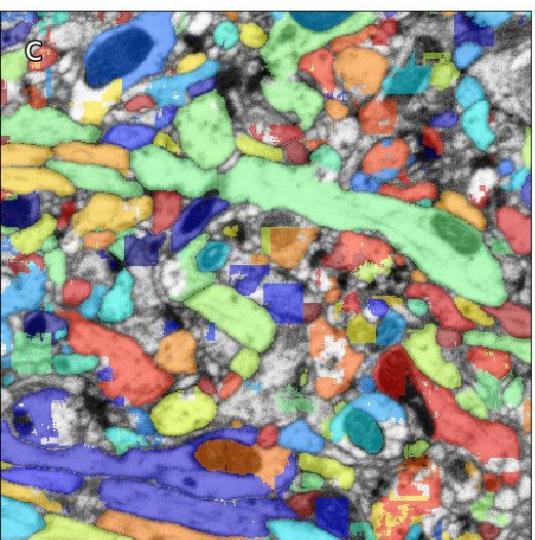
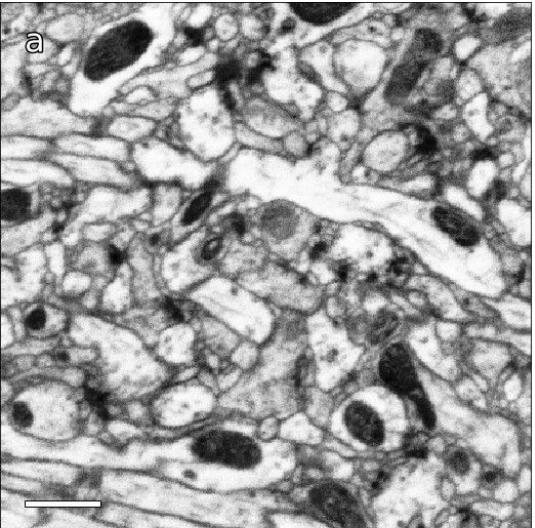
OpenNEURO

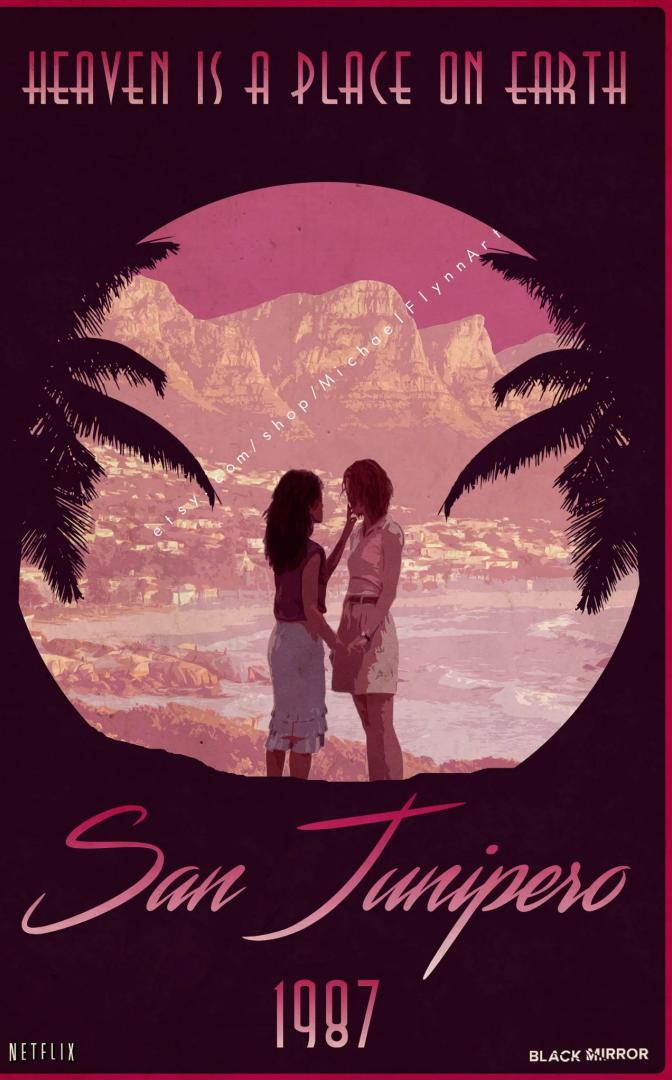


what you want is the
inner thought [REDACTED]
monologue of your
brain [...] if we had a
billion hours of that,
AGI is here roughly
speaking [REDACTED]

Andrej Karpathy







A fast path toward safe brain-like AGI

- Get brain data under high-entropy naturalistic conditions
- Build a zoo of virtual bodies and environments
- Leverage existing data archives to build foundation models
 - Bring in closed data
- Accelerate cost-cutting data acquisition technologies, e.g. in connectomics and wireless free-behavior electrophysiology

NeuroAI for AI safety

Niccolò Zanichelli¹

Joanne Zichen Peng¹

Patrick Mineault¹

¹Amaranth Foundation

✉ patrick@amaranth.foundation

𝕏 @patrickmineault

🔗 neuroai.science

