Cause Area Analysis: Differential Neurotechnology Development and Governance

In three sentences

Neurotechnology is being developed today that could have extremely positive or negative impacts on the wellbeing of humanity in the near-term and long-term future.

Almost no effort is being made to proactively steer neurotechnology development toward good outcomes.

There are fundable projects today that can improve the likelihood of good outcomes.

Summary

Importance
Neurotechnology could have extremely positive or negative impacts on the wellbeing of humanity and other beings in the near-term and long-term future.

In the positive direction, neurotechnology is needed to address a growing ~21% share of global disease burden, has the potential to eliminate vast amounts of unnecessary suffering, and may aid in the development of safe advanced AI.

In the negative direction, it might facilitate totalitarianism or irreparably corrupt human values.

Influencing the development of neurotechnology may be quite urgent. Without governance, neurotechnologies currently in-development could, within the next 5 or 10 years, become locked into a trajectory toward negative outcomes that will be hard to alter. With sufficient effort, neurotechnology that could benefit pressing concerns like AI safety could realistically be developed within 10 to 20 years.

Neglectedness
While neuroscience (the study of nervous systems in general) receives ~$20B/year in funding globally, efforts toward differential neurotechnology development or governance are limited to small amounts of academic research, government rhetoric, and the stated missions of a few companies.

Tractability

There is an opportunity to shape the field of neurotechnology from the beginning, in a similar spirit to early action on AI governance, space policy, and biosecurity.

A new philanthropist could exert considerable influence over the future development of neurotechnology by establishing governance mechanisms like patent pools, acquiring key intellectual property, strategically building and controlling R&D infrastructure, and directly funding differential development of beneficial neurotechnologies.

Table of Contents

Importance

A neurotechnology is any tool that exogenously observes or manipulates the state of living biological nervous systems, especially the central nervous system of humans.[1] Readers may be familiar with neurotechnologies like electrode-based brain-computer interfaces (BCIs), antidepressant drugs, MRI, transcranial magnetic stimulation, or optogenetics.

Neurotechnology could have extremely positive or negative impacts on the welfare of humanity and other beings in the near- and long-term future.

Influencing the development of neurotechnology may be quite urgent, because neurotechnologies currently in development could strongly benefit humanity's pressing AI safety and other existential-risk-reduction efforts if accelerated, and because preventing negative impacts of neurotechnology may be easier with early intervention.

The Potential Impacts of Neurotechnology

This section lays out some potential impacts of neurotechnology without considering the timeframes in which they may be achievable. We discuss timelines further below.

---

[1] This definition includes not just electromagnetic brain-computer interfaces (BCIs) but also chemical, biological, mechanical, and other modalities. Activities like exercise or media consumption are not neurotechnologies, since their effects on cognition are mediated through endogenous pathways. Meditation, hypnosis, and other modalities that influence brain function exclusively via unusual forms of conscious engagement could be considered neurotechnologies, but we won't consider them here.

Treating neurological and neuropsychiatric disorders

Most R&D in neurotechnology today is focused on treating neurological and neuropsychiatric disorders. The two are different,[2] but for simplicity in this section we're going to combine them under the name "neuro disorders".

Based on data from the Institute for Health Metrics and Evaluation in 2019 (database, publication, our calculations), **neuro disorders account for ~21% of global disease burden**. This is ~530M DALYs (disability-adjusted life years) out of ~2.54B total. Using Open Philanthropy's estimate of $100k USD/DALY (source, section 3.4), the value of having provided cures for all neuro disorders in 2019 would have been $53 trillion. This is an overestimate in that people kept alive by cures for neuro disorders would have been afflicted by other diseases, but it also does not account for the value to future people spared from those disorders.

For comparison, ischemic heart disease alone accounted for ~7% of global disease burden in 2019 and communicable, maternal, neonatal, and nutritional diseases combined accounted for ~26%.

The percentage of global disease burden caused by neuro disorders has steadily increased from 16% in 1999. It will presumably continue to increase as communicable disease treatment and maternal health improve globally. In high socio-demographic-index countries, neuro disorders accounted for ~30% of disease burden in 2019, barely up from ~29% in 1999.

We have not thoroughly vetted the methods used to obtain the IHME's data, and the IHME came under strong criticism for its poor COVID-19 modeling. As some corroboration, the WHO's global disease burden estimates roughly match (±5% of top-line numbers) the IHME's estimates. The WHO's estimates include the IHME's data as one source, but also claims to include data from national health registries, WHO technical programmes, United Nations partners, and other scientific studies.

The global burden of neuropsychiatric disease may be significantly underestimated, as The Happier Lives Institute suggests in a recent report. (source, section 3.2) Reasons for underestimation include that suicide and self-harm are not counted as neuropsychiatric disease burden (though we count them in our estimates above), that self-report and diagnosis of emotional symptoms is lower in non-Western countries, and that disability weights for mental disorders are underestimated.

The cost-effectiveness of developing neurotechnology to treat and cure neuro disorders is difficult to estimate since no particular neurotechnology is guaranteed to cure any particular neuro disorder. In addition, it is beyond the scope of this document to estimate how much the burden of disease would be reduced simply by improving access to the current best treatment

[2] The line between neurological and neuropsychiatric disorders is fuzzy. The former generally refers to diseases like Alzheimer's or Parkinson's with observable pathologies to the structure or activity of neurons, while the latter includes diseases like depression and ADHD that don't (yet) have understood mechanisms.

options or through public health initiatives. The Happier Lives Institute report on mental health delves into much greater depth on the topic of cost-effectiveness of treating neuropsychiatric disorders in lower-income populations.

But it should be emphasized that current treatment options for neuro disorders are quite poor compared to those for disease areas like infectious disease or neonatal disorders. At present there are no curative treatments for any neurological diseases, only ways to manage symptoms to greater or lesser degree. (source) There are also no reliably curative treatments for most neuropsychiatric disorders. (source) New neurotechnologies are required to significantly improve treatment outcomes.

Direct manipulation of subjective wellbeing

The subjective wellbeing of a conscious organism is, to the best of our knowledge, exclusively a function of the physiological state of its brain. **Ultimately, neurotechnological manipulation is the only feasible means of alleviating all unnecessary and unproductive suffering and maximizing subjective wellbeing.**[3]

It is also potentially a source of tremendous suffering.

Much suffering can be alleviated by changing an organism's external circumstances. But subjective wellbeing is far from perfectly correlated with external circumstances. Depression, other mood disorders, and chronic pain are evidence of this, as is hyperthymia in the happier direction.

Beyond named neuro disorders, which we've addressed in the previous section, lived experience suggests that most of us spend much of our lives experiencing unnecessary and unproductive suffering, however intermittently. (This is not to say that all suffering is unnecessary or unproductive.) The experience of non-human animals may be similar or considerably worse. Neurotechnology is the only way such suffering can ever be fully addressed.

On the positive end, we have no idea how good lived experience can get. The upper limits of how much subjective wellbeing is possible to achieve via direct manipulation of brain states is unknown. What we consider a good life today may be considered torture by the standards of a society with adequate neurotechnology.

Trying to quantify any of the claims above is an exercise in false precision given the crudeness of measures of subjective well being. Measures like DALYs and QALYs (quality-adjusted life years) are not designed to account for changes in subjective well-being. For example, "an intervention that made mentally healthy people happier would avert zero DALYs." Measures that take subjective well-being into account, like "Well-being adjusted life years" (WELLBYs) or other estimates based on self-reports have the benefit of being direct, but face calibration challenges.

[3] Whether subjective wellbeing and wellbeing are equivalent is beyond the scope of this document.

For example, reporting life satisfaction on a 0-10 scale assumes life satisfaction is a bounded quantity. This is not to impugn these attempts at measurement - they are worthy attempts at an important problem. But for now we consider measuring subjective wellbeing an open problem.

Unfortunately, neurotechnology capable of maximizing wellbeing may well be capable of maximizing suffering.

Substance abuse is the most familiar means by which neurotechnologies cause suffering in the modern world. We are not aware of rigorous estimates of the subjective suffering caused by substance abuse, but estimates of the economic impacts range from $10s to $100s of billions per year in the U.S. depending on which factors are included (value of lives lost, lost work productivity, health care costs, crime, etc.). (1, 2, 3, 4) Substance abuse is an example of the more general concept of *wireheading*: using neurotechnology to directly manipulate pleasure or motivation systems in the brain, usually in a way that is overall harmful.

Beyond this, sufficiently advanced neurotechnology is also an opportunity for horrific abuses, such as making a victim feel intense suffering while masking outward signs of suffering. Just as we do not know how good subjective wellbeing can be, we also do not know how bad it can be.

Enhancement and value shift

Neurotechnology may offer many ways to enhance human abilities:

- Improved control of memory formation/erasure/reconsolidation, including accelerated learning
- Improved access to information with brain-computer interfaces
- Improved manipulation of devices and tools with brain-computer interfaces
- Improved concentration
- Control of energy level
- Control of emotions, including in decision-making
- Improved impulse control
- Improved introspection
- Altering personality traits
- Flagging or eliminating cognitive biases (future discounting, status quo bias, etc.)

Some more speculative enhancements — including those perhaps better described as new abilities rather than enhancements — are listed here.

Many people would benefit from cognitive, behavioral, or emotional enhancement on the margin in their personal lives or careers. And a general increase in wisdom and rationality might be useful for improving and safeguarding humanity. (Source, Security and Stability section) These are central motivations for the Rationality movement with initiatives like The Center for Applied Rationality and for the interest in community epistemic health in the Effective Altruism movement.

Neurotechnology offers much stronger potential to improve individual reasoning and cognition than has been previously examined by Effective Altruist organizations. More research is warranted on how such enhancement would affect areas of interest to Open Philanthropy, such as mitigating risks from great power conflict, global and space governance, and expanding the number of effective altruists.

But neurotechnologies are not the only means of achieving some of these enhancements. And they risk causing native human capabilities to atrophy, as arguably calculators have done to our ability for mental arithmetic. Worse, they might provide malicious actors opportunities to manipulate people's bodies or thoughts, though security will doubtless be a top priority for any neurotechnology developer.

Most concerningly, though, many desirable neurotechnological enhancements achieve their effects by manipulating a user's beliefs and motivation systems, which are tied in complex ways to their goals and values. Thus **value shift is a risk with the adoption of any neurotechnology.**

These value shifts could be accidental or malicious. For example, a neurotechnology that increases empathy (of which several are in clinical trials) could lead to users allowing themselves to be taken advantage of or make society lenient in dealing with psychopaths or despots. Or a totalitarian government could also use neurotechnologies to monitor change in the beliefs, goals, or values of its citizens. (Caplan 2008, A New X-Risk Factor: Brain-Computer Interfaces) For example, one could imagine a reeducation camp that actually reeducated people with 99% success, or what would have happened if MKUltra had achieved its goals. Between these two extremes there is a question of the morality of using neurotechnology to reform criminal behavior, perhaps as an alternative to incarceration.

How should the legal system - or we as individuals, for that matter - treat a person who accidentally changes their motivations to ones that they *a priori* would not have wanted, but *a posteriori* want to maintain? How can we distinguish between persuasion and coercion in a world where neurotechnology permits a continuum of communication forms between speech and thought?
(TODO look into this and other literature on the topic. There must be more recent work. Philosophy of 2nd order preferences or something? There is certainly fiction on it, e.g. Greg Egan's Quarantine.)
TODO add Wei Dai on this

Governance, public education, and shaping the field of neurotechnology early may all help ensure good outcomes. But it will be challenging to resist value shifts caused by adoption of neurotechnologies that improve people's earning potential or societal status.

Consciousness and welfarism

While conscious experience may or may not be a determinant of moral patienthood, **there are many questions about consciousness whose answers would strongly affect welfarist**

reasoning. Some of these questions are raised in Open Philanthropy's 2017 Report on Consciousness and Moral Patienthood.

Pivotal questions include:

- What are the neural correlates of suffering? Can we identify them in non-humans? How much of our dreams are spent suffering, and is this morally relevant? Are certain short- or long-term memory processes required for suffering?
    - Neuroimaging technology may help identify correlates, and control over memory processes may help resolve the latter question.
- To what degree is consciousness substrate-independent?
    - Neurotechnology could provide evidence relevant to this question via “partial uploading” experiments, wherein a part of the brain is anesthetized and researchers attempt to mimic its function by exogenous stimulation such that an awake subject can’t tell the difference. The utility elevance of such experiments has been debated.
- More generally, what is the map of the landscape of conscious experience? Are concepts like hedonic valence, emotions, or moods reliable axes of variation of conscious states? What are their neural correlates? Is there a continuum of “more” or “less” consciousness?
    - The finer degree of control neurotechnology gives us over neural activity, the better we can answer this question. The better a description we can achieve, the better we may be able to define and measure subjective wellbeing and assess non-health, non-pecuniary benefits.
- How many independent consciousnesses exist in a human brain? Do they make each other suffer?
    - Neurotechnology could be used to isolate areas of the brain from interaction with other areas and communicate with them independently, as was potentially done historically during callosotomy operations. Or it could be used to instrument areas of the brain that cannot normally communicate with means of doing so.
- How continuous in time is conscious experience? What are the shortest and longest intervals of conscious experience? Do different brains — or different parts of the same brain — run at different "clock speeds"?
    - TODO better operationalization of David Eagleman’s experiments? Can we increase the flicker-fusion rate of humans?
- Is consciousness necessary for moral patienthood?
    - This isn’t an empirical question, but findings made using neurotechnology may be relevant to our beliefs about it. For example, can we use neurotechnology to

induce “p-zombie-like” states? Mimicking the neural processes implicated in sleepwalking or bipolar blackouts might enable subjects to exhibit phenomena we associate with moral patients — like engaging in conversation — despite them later reporting themselves as unconscious.

Some of these questions may turn out to be ill-posed or irresolvable by scientific inquiry. **But in general, understanding welfare requires better neuroscience, and better neuroscience requires better neurotechnology.**

We won’t attempt here to estimate a dollar value of resolving any of the above questions, but the values could conceivably be extremely large. E.g. suppose a series of neuroscientific results drastically increased our estimate of the amount of wild animal suffering.

Impacts on AI Safety

Neurotechnology may have positive impacts on the development of safe AI. The interplay between AI and neurotechnology has been discussed previously (niplav, Long, Bensinger, Byrnes), but remains underexplored. **Because neurotechnology mostly does not compete for talent with other areas of AI safety research, it is worth investigating as part of an “all hands on deck” approach to AI safety.**

Neurotechnology and Outer Alignment

Outer alignment is the task of giving an optimizer (i.e. an AI system) an objective function whose solutions are those the designer intended. Usually the desired solution is something like “act according to human values”.

Getting more data on human values

Outer alignment of AIs to human values may not be a well-posed task inasmuch as “human values” may not be well-defined. But even if they are, it is not clear from what evidence an AI could or should infer them. **Neurotechnology could provide greater quantity and quality of data on human values to improve outer alignment.** This may be partly what Neuralink’s mission of “merging humans with AI” refers to.

Moral judgments are one source of data about human values. A number of proposals for building safe AI rely on access to this kind of human feedback. Typically these judgements are obtained via language or other consciously expressed feedback like voting. Neuroimaging technology could facilitate access to these judgments and increase the amount of data available for AIs to learn from. This could be done passively throughout daily life combined with e.g. smart glasses to record the situation in which the moral judgment is being rendered. Neuroimaging could also increase the quality of moral judgements obtained by disentangling them from corrupting processes like motivated reasoning and memory reconsolidation.

Subjective wellbeing is another source of data about human values. While increasing subjective wellbeing is not a primary goal of many moral systems, it is central to many. Neuroimaging

technology could drastically increase our ability to measure and track it, which is important given how poorly we predict our future wellbeing. For example, the IHME's DALY estimates suggest people estimate moderate depression to be about 2-4x worse than living with a limp. (source, mild or moderate major depressive disorder vs. conditions with limp as a symptom) However, people who have really suffered from both mobility issues and depression report that depression is 10x worse for their well-being. (source, section 4.4, derived from Table 2 here)

Human interpretability

AI interpretability is valuable because the input/output behavior of an AI on a single dataset is insufficient for us to know what its behavior will be in general. Humans are similar: our input/output behavior (what be called revealed preferences) is insufficient to infer our values.

Obtaining a mechanistic, causal understanding of the neural processes that underlie human action and moral judgements is ultimately necessary to understand and operationalize what individual human values are. Neurotechnology is needed to obtain this understanding. Put another way: **ultimately human interpretability is as important as AI interpretability,** though for different reasons and perhaps on different timelines (e.g. AI interpretability tools may be more important in the near-term to check for obviously dangerous motivations, like the desire to break out of a virtual machine). Human interpretability helps specify human values; AI interpretability helps determine whether an AI is hewing to them.

Dealing with intersubjectivity

It is an open question whether "human values" refer to any individual's values at all, or whether in fact human values refer to intersubjective values that are distinct from, and often opposed to, individual values.[4]

Neurotechnology might help with this problem by providing sufficient data to train models that reproduce individual human value judgments. This can be thought of as partial whole-brain emulation. (Gwern) It is conceivable that these models could predict the moral judgements of an individual with error significantly less than the variance in judgements between human beings. This is quite speculative, but a sufficient number of such partial emulations could serve as a "moral parliament" to an advanced AI system.

Neurotechnology and Inner Alignment

Inner alignment is the task of designing an optimizer (i.e. an AI system) that finds a correct solution to the (inevitably underspecified) outer optimization problem we want it to solve.

Given that the human brain may be the best available example of an aligned intelligence, emulating its operation may prove useful for designing inner-aligned systems. **Doing so will require a greater understanding of neuroscience than we have now, and better neuroscience requires better neurotechnology.**

[4] Though since individual human brains will be the ones deciding whether human values are intersubjective, the problem may just reduce to "get close to the value function of individual humans," even if those value functions prefer to disregard themselves as the ultimate source of value.

For example, it might be the case that AI systems whose architecture mimics the optimization system of human brains will be more inner-aligned. (Byrnes, Jilk, et al.) In the limit of perfect mimicry one would achieve whole-brain emulation, which is aligned by definition, but much less perfect mimicry might still yield more inner-aligned systems. Neurotechnology is necessary to understand how the brain's optimization system works, which we do not currently understand.

Or perhaps human agentic behavior can be disentangled into sub-behaviors, and AI systems can be built to perform the safer sub-behaviors. E.g. perhaps different neural circuits control exploration and self-preservation behaviors.

It might also be possible to build "tool AI": systems based on non-optimizing, non-agentic aspects of neural computation. AI systems designed based on these features of the human brain could be valuable to humanity but less risky than agentic AI systems, despite being less competitive.

Or it might be the case that hybrid human-AI systems can be built where key decision-making or goal-setting parts of the architecture are delegated to circuitry in real human brains. Hybrid systems are typically assumed to not be competitive with pure AI systems in the long run, but they may be useful during particular stages of AI development.

Human Enhancement and AI Safety

Even in a world where safe AI is developed, it only takes one defector building an unsafe AI to cause bad outcomes. **Neurotechnology may offer ways of enhancing coordination**.

For example, if high-accuracy lie detection was developed, companies in control of AGI-enabling hardware could choose to only sell to customers who neurotechnologically verified their commitment to not building certain risky AI technologies. Even stricter means of coordination enforcement might be possible with neurotechnology that can monitor or modify behavior. While such solutions may sound draconian, they do not require coercion.

And as mentioned above, movements like the Rationality movement are predicated on the idea that improving human reasoning ability would be beneficial for human flourishing, perhaps including the ability to understand and perform well in coordination problems. Neurotechnology is the most promising means of significant, large-scale improvements in individual human rationality.

Risks and uncertainties

Different strategies for AI safety carry different risks, and neurotechnological ones are no exception.

For example, neurotechnology might offer AI systems an additional "attack surface" by which to influence human judgment and values. Technologies that can only sense but not manipulate neural activity might mitigate most of this attack surface while still being useful for outer

alignment. But even pure-sensing neuroimaging technologies could give a malicious AI system a clearer measure of whether it successfully altered human values by persuasion or other means. Then again, it might be that an AI would have to be superintelligent to manipulate human values in this way, at which point the existence of neurotechnology is irrelevant: the AI will be perfectly capable of harming humanity in more straightforward ways.

Another risk is that greater understanding of neuroscience could be an accelerant to developing transformative AI systems by yielding new algorithmic ideas, like how current state-of-the-art deep learning systems were loosely inspired by integrate-and-fire neuron models and other ideas from neuroscience.

Much more research is warranted to assess the risk-benefit tradeoff of specific neurotechnology use cases and compare them to the risks of other AI safety strategies.

Urgency

Rather than make forecasts for each specific impact described in the previous section (though we think this would be worthwhile), we will review emerging neurotechnologies and consider what capabilities they might afford on what timelines.

Our tentative conclusion from what follows is that **influencing the development of neurotechnology may be quite urgent**. Without intervention, neurotechnologies that are currently in preclinical and clinical development could, within the next 5 or 10 years, become locked into a trajectory toward negative outcomes that will be hard to alter. They could also, with sufficient effort, be developed within 10 to 20 years to the point that they would meaningfully benefit pressing concerns like AI safety (in addition to other, potentially less-urgent benefits).

What neurotechnologies exist or are in development?

The following subsections summarize the current state of neurotechnology R&D, with neurotechnologies grouped by their stage of maturity, from most mature to least.

Sidebar: key characteristics of neurotechnologies
Neurotechnologies can be characterized along many dimensions, but a few key, high-level factors we will mention are:

- Sensing vs. manipulating: how much the neurotechnology reads vs. writes the brain
- Spatial resolution: how finely in space the neurotechnology can sense or manipulate neural tissue
- Spatial extent: how much of the brain or nervous system it can access
- Temporal resolution: with what frequency it can sense or manipulate neural activity
- Substrate specificity: what biological material(s) it senses or acts on
- Flexibility: how easy is it to reprogram or alter the behavior of the neurotechnology during or between uses

- Safety: what risks the user faces
- User ease: how easy it is to adopt, wear, implant, use, or maintain

Neurotechnologies that are currently FDA-approved or widely used

- Small molecule drugs (too many to name):
    - Examples:
        - Stimulants (caffeine, Adderall)
        - Antidepressants (Prozac, Wellbutrin)
        - Anesthetics (morphine)
        - Anxiolytics/sedatives (Xanax, Valium)
        - Psychedelics (LSD, psilocybin)
        - Empathogens/entactogens (MDMA)
- EEG
- ECoG
- (f)MRI
- Transcranial magnetic stimulation (approved for treatment-resistant depression, anxiety, OCD, and smoking cessation)
- Electroconvulsive therapy (approved for depression and too many others to name)
- Deep brain stimulation (approved for Parkinson's, movement disorders, and OCD)
    - Medtronic
- Peripheral nerve stimulation
    - Vagus nerve stimulation (approved for epilepsy and depression)
        - E.g. LivaNova
    - External stimulators for tremor
        - E.g. Cala Health
- Surgical tools (too many to name, but important ones include):
    - Neurovascular stents
    - Stereotactic surgical equipment
    - Skull reconstruction implants
- Cochlear implants
    - Cochlear
    - Oticon
- Retinal implants
    - E.g. Second Sight (humanitarian use, now defunct)

Neurotechnologies currently in or enrolling human clinical trials

- Subdural motor BCI
    - Blackrock Neurotech - suite of products for neuroscience research projects (electrodes, data acquisition systems, headstage).
    - Paradromics - collects neural signals with a fully implantable device to address medical challenges. Microelectrode arrays target neurons 1.5 mm below the surface of the cortical brain.
    - Neuralink - designing the first neural implant that will let you control a computer or mobile device. Micron-scale threads are inserted into the motor cortex.

- Endoscopic motor BCI
  - Synchron - metallic mesh tube with electrode contacts placed inside a blood vessel in the motor cortex. Does not require brain surgery.
- Peripheral BCI
  - BrainRobotics - multichannel electromyography sensors in the wrist, which enable a prosthetic hand to process muscle signals from the user's arm.
  - CTRL-labs
- Cortical stimulation for memory enhancement
  - Nia Therapeutics - precision brain stimulation therapies to treat memory loss due to brain injury and degenerative disease.
  - Braingrade
- Retinal implants
  - Pixium - intended to partially replace the normal physiological function of the eye's photoreceptor cells by electrically stimulating the nerve cells of the inner retina.
- Functional ultrasound neuroimaging
  - Iconeus - system for imaging changes in cerebral blood volume. This makes it possible to follow changes in neuronal activation over time. Currently used in animal research studies.
- Functional photoacoustic neuroimaging
  - Massively parallel functional photoacoustic computed tomography of the human brain
- Transcranial electrical stimulation
  - Temporal interference stimulation - non-invasive brain stimulation. Can activate neurons in target brain regions with high-frequency carriers.
- Transcranial ultrasound stimulation
  - Brainsonix - target specific neuronal circuits using fMRI and repair the circuits by activating or inhibiting them.
  - others, including several stealth companies
- Ultrasound-mediated BBB opening
  - Carthera (Sonocloudultrasound to temporarily disrupt the blood-brain barrier (BBB) enabling a window where drug therapies can be administered.
- fNIRS)
  - Kernel Flow - brain measurement systems using TD-fNIRS
  - (TODO add the methods from Maria Franceschini's BRAIN talk that are in or have been in clinical trials)
- DCS - Diffuse correlation spectroscopy. Started human trials with pulsed laser at at 1064nm.
- Peripheral nerve stimulation
  - Audiovisual stimulation
    - Cognito - non-invasive neuromodulation with the potential to improve outcomes in a range of neurodegenerative diseases, including Alzheimer's disease.
  - Vagus nerve stimulation

        - Sharper Sense - stimulates Vagus nerve for improved sensory processing.
        - Setpoint - Therapy for Biologic-Refractory Rheumatoid Arthritis.
    - Splenic nerve stimulation
        - Galvani - stimulates splenic nerve to treat Rheumatoid arthritis (RA).
    - Vestibular nerve stimulation
        - Neurovalens - transdermal activation of the homeostatic nuclei of the brainstem and hypothalamus, allowing for alterations in autonomic function, circadian regulation and Neuro-metabolic influence.
- Spinal cord stimulation
    - Onward - enable people with spinal cord injury to move again, aided by programmed stimulation of the spinal cord.
- Gene therapy
    - AAV-delivered genes for numerous neurological diseases
- Cell therapy
    - Cell therapies for stroke and Parkinson's
- Too many drug candidates to name
    - Psychedelic therapeutics are experiencing an unusually fast pace of development due to recent regulatory and social changes in the United States

Neurotechnologies in preclinical development

Preclinical development means a technology has not yet been (to our knowledge) deployed in humans.

- Next-generation subdural BCI
    - Precision Neuroscience - thin-film microelectrodes designed for rapid, minimally invasive deployment on the cortical surface.
    - Integrated neurophotonics - implanting an entire lens-less imaging system within the brain itself by distributing dense arrays of microscale photonic emitter and detector pixels.
- Endoscopic stimulation
    - ME-BIT - in vivo proof-of-concept testing of an endovascular wireless and battery-free millimetric implant for the stimulation of specific peripheral nerves.
- Distributed implanted stimulators (“neural dust”)
    - Iota - millimeter-sized, ultrasonic-powered bioelectronic “neural dust” built to interface directly with the central nervous system.
- Peripheral BCI
    - Science.xyz - “all of the information that flows in or out goes through only a handful of nerves in the head and spine. These form the complete “API” of the body: if you can connect to them with single-unit resolution, you can provide exactly the same senses and motor surface that your nose, eyes, ears, hands and so on do. (source)
- Next-gen fNIRS

    - Openwater - emits laser light detected by camera chip and decoded to get the speed of blood flow.
    - CoMind
- MEG
    - Sonera
    - Kernel Flux
- Gene therapy
    - US-mediated viral delivery - needle-free combination of focused ultrasound-mediated viral delivery and extracorporeal illumination with red light, to achieve selective neuronal activation at depths up to 4 mm.
    - Non-AAV gene delivery vectors, e.g. Ensoma
    - Optogenetics - stimulation strategy based on the expression of light-sensitive proteins in the neuronal cell membrane that, upon illumination, alter the electric state of the neuron.
    - Sonogenetics - allow ultrasound to connect directly to cellular functions such as gene expression.
    - Sonomagnetics- sono magnetic stimulation (SMS), that can generate an electrical current focused in a small volume deep in neural tissue.
- Cell therapy
    - Optogenetically engineered "living electrodes"
- Too many drug candidates to name

Outside view on development timelines

As a prior, 20 years has been given as a rough estimate for how long it takes an invention to translate into an adopted technology. It also appears that the pace of adoption of new technologies is increasing as time goes on, but we have not seen reliable statistics for this fact.

Reference class estimate

To improve on this estimate, here are summaries of the development of several neurotechnologies and related technologies:

Deep Brain Stimulators (source)

- Building on extant stereotactic neurosurgical tools and cardiac pacemaker technology, prototype DBS systems were first implanted in humans in the late 1960s.
- Implanted in numerous patients until 1976, when the FDA is established. They stop DBS sales until clinical trial data is submitted.
- No company is willing to run trials until the neurology field establishes clearer standards for patient improvement.
- Once they do, in 1997 Medtronic runs trials and gets FDA approval for essential tremor and some Parkinson's cases.
- FDA approval for all Parkinson's cases in 2002 after more trials.
- 40k individuals treated with DBS within 10 years of approval.

Summary: ~40 years from first human use to consistent human use, including ~20 year pause to convince FDA.

Cochlear Implants (source)

- First implantation of electrodes to explore restoration of hearing loss in 1957.
- By 1977 twenty-two patients had prototype implants.
- FDA approval for adults in 1984.
- Slow adoption because the adult deaf community was generally not interested in, and sometimes hostile to, the idea of becoming hearing people.
- Pediatric cochlear implants were approved in 1990, where there was stronger uptake. (90% of deaf children have hearing parents.)
- By 2009 there had been in the 100ks of implants total. This may be only 10% of the total addressable market. (source)

Summary: on the order of 50 years from first human implant to consistent human use, but ~15 years from FDA approval in a market with demand

Transcranial Magnetic Stimulation (source)

- First demonstration of magnetic stimulation in 1896
- Single-pulse system demonstrated in humans in 1985
- Repeated-pulse system developed and effects on depression reported by 1994
- FDA approval for depression treatment in 2008. Arguably this would have gone faster had IP been handled better

Summary: ~9 years to development basic invention into therapeutic system, ~12 years to get approved, widely used today but still a small fraction of neuropsychiatric treatments

Transcranial Electrical Stimulation (source)

- People have been running electricity through their heads since antiquity, including FDA approvals for electroconvulsive therapy and devices for treating migraine
- Two papers around 1998 reignited interest in low-output (<10 mA) transcranial electrical stimulation for modifying cortical excitability
- By 2006 a few articles about these systems make it into newspapers
- By 2012 DIY kits are being sold on the internet
- By 2014 startups like Halo and Thync have been started
- No FDA approvals have been made for low-output systems to date

Summary: ~6 years from popularization to DIY systems, with startups following immediately after

Stentrode

- Building on extant neurovascular stent technology, Synchron (originally named SmartStent) was founded in 2012 and developed their stent-based BCI prototype with funding from DARPA, and others. (source)
- First publication in 2016 demonstrating Stentrode in sheep. (source)
- Synchron got IDE approval for clinical trials from the FDA in 2021 and performed their first human implantation in 2022.

Summary: ~6 years from conception to (published) sheep and ~6 years from sheep to first human.

Prozac (fluoxetine) (source)

- First synthesized at Lilly in 1972
- FDA approved for depression in 1987, the first SSRI to be marketed
- Hailed as a breakthrough, eventually became 1/4 of Lilly's revenue, >40M patients received it by 2002. (Many more had taken other SSRIs.)
- Consistently in the top 30 most-prescribed drugs in the U.S. by one estimate

Summary: 15 years from synthesis to approval, followed by widespread adoption almost immediately

LSD (source)

- First synthesized in 1938.
- First ingested in 1943. (source)
- Sandoz started marketing the drug in 1947 for a variety of uses.
- The CIA reportedly bought the world's entire supply in the early 1950's for use in the MK-ULTRA program. (source)
- Became popular recreationally from 1960s onward.
- Made illegal in U.S. in 1968. (source)
- Recently use has reportedly increased, and has been decriminalized in one state.
- An estimated ~10% people in the U.S. have used LSD in their lifetime. (source) Similar rates are reported for Australia. (source)

Summary: ~5 years from synthesis to discovery of effects, ~15 years until popular use began, despite tortuous history remains widely used

Mobile phones

- First mobile phone demonstrated in 1973. (source)
- First commercial offering 1983. (source)
- Usage in U.S. households rose from 10% in 1994 to 63% in 2004. (source)

Summary: ~10 years from prototype to commercial product, ~20 more years to ubiquity, with a significant inflection

Personal computers (source)

- Xerox Alto demonstrated in 1973
- Apple Macintosh released in 1984
- Usage in U.S. households rose from 20% in 1992 to 63% in 2003

Summary: ~10 years from prototype to mass commercial product, ~10 years to ubiquity

Breast augmentation (source)

- First breast implant surgery in 1962.
- FDA bans silicone implants in 1992, saline implants become dominant. (source)
- ~100k breast augmentation surgeries in 1997. (source)
- ~300k breast augmentation surgeries in 2018 and 2019

- An estimated 4% of women in the U.S. have had breast augmentation as of 2014.

Summary: ~30 years to becoming a standard procedure, with fairly linear growth

LASIK eye surgery (source)

- Building on knowledge from existing non-laser keratotomy surgeries, LASIK was conceived in 1988. (Similar procedures were being developed concurrently.)
- First LASIK surgery performed in U.S. in 1992.
- FDA approved devices for LASIK in 1998.
- Adoption rapidly increased to ~1.2M surgeries per year in the 2000s, then tapered to ~700k/yr in the 2010s (source)

Summary: ~4 years from conception to demonstration in humans, ~8 more years to become a standard procedure

The key events in the timelines above are not directly comparable and vary in time from 1 to 4 decades from conception to widespread use. But we can say that (1) none went from conception to widespread use in less than 10 years and (2) the 20 year prior stated above from conception to widespread adoption seems short. An estimate of 30 years from conception to widespread adoption seems more reasonable, acknowledging that this has large variance. And a conservative estimate of 15 years from prototype demonstration (in humans, where relevant) to widespread adoption, loosely defined, also seems reasonable.

Expert surveys and forecasts

We could only find surveys for BCI technology rather than neurotechnology as a whole, and both were too nonspecific to extract meaningful information from. (source 1, source 2)

Inside view on development timelines

The following are key factors that can influence the development timelines of a neurotechnology.

- **Desirability of effects**: How much utility does the neurotechnology provide, and how easy is it to use?
    - Historically, potency of a neurotechnology has been in tension with regulatory burden: the more potent, the more regulated. This is especially true of neurotechnologies that are pleasurable to use.
    - Noninvasiveness and reversibility of a neurotechnology is often in tension with ease of use. E.g. an implant is burdensome to get, but in the long run may be preferable to wearing a headset.
- **Market size**: how many users will the neurotechnology have?
    - Neurotechnologies treating specific medical indications have the advantage of a nearly-guaranteed financial payoff (medical insurance reimbursement, mostly in the U.S.) and known number of potential customers. Neurotechnologies for consumers may have larger markets and currently face less regulatory burden than medical devices, but with much more uncertainty about adoption.

    - Even if every adult with a diagnosed mental health disorder received a neural implant within 10 years, that would still only be ~25% of the population.
- **Regulation**: How much regulatory burden does the neurotechnology face?
    - Drug regulation
        - In the U.S. chemical and biologic neurotechnologies are regulated by the FDA for medical uses and by the DEA in general.
        - Anyone can create a new chemical or biologic with any effect and sell it in the U.S. without interference from the FDA, provided they make no claims about it treating or curing any disease. But if it has potent enough effects, the DEA will likely exert control over its distribution.
    - Medical device regulation
        - In the U.S. the FDA decides which neurotechnologies count as medical devices and what degree of clinical evidence they require to be marketed.
        - European medical device regulation is considered less burdensome.
    - Surgical regulation
        - In the U.S. surgical procedures are not regulated directly at the federal level. States sometimes have laws around specific procedures like abortion or cosmetic body modifications
        - Surgeons can have their licenses revoked for performing operations outside their scope of practice.
    - Consumer goods regulation
        - In the U.S. the FTC, which takes action against "hazardous…products…without adequate disclosures"
        - The CPSC, which takes action against "unreasonable risks of injury", though not in the FDA's remit.
        - The FCC, which regulates devices that emit RF signals.
- **Market power**: how much ability do single actors have to manipulate the direction of the field?
    - Unlike in the software industry, intellectual property affords single actors significant control over the availability of neurotechnologies.
    - Several large companies and university technology-transfer offices are frequent "bad actors" in the neurotechnology space at present, stifling competition and new market entrants. This behavior may increase if the neurotechnology market grows.
- **Cultural resistance**: beyond simply not attracting many users, will the neurotechnology face active resistance from the public?
    - Cultural resistance can lead to regulation or influence government, as with e.g. the U.S. War on Drugs or the societal pushback on the BrainCo headband that was piloted to increase focus in Chinese schools.
    - It can also increase adoption or divide groups, as with
- **Iteration speed in humans**: how quickly can new advances be designed, built, and tested?
    - In general, the less time, money, and effort it takes to develop each new iteration of a technology, the faster the technology will improve.

    - A neurotechnology that allows an app-store- or home-brewing-like degree of end-user customization and open market development of new features will yield new capabilities much faster than one treated as a medical device, to which all changes must be re-approved by a regulatory agency and justified with clinical data
    - Faster iteration times yield more serendipitous discovery and capabilities development, but also poses safety concerns.
- **Extrinsic shocks:** how will events and trends outside the field of neurotechnology influence it?
    - If humanity is decimated by a global nuclear war or pandemic, neurotechnology development will proceed rather slowly.

Conclusions

Development timelines in the absence of intervention

Neurotechnologies that are currently in preclinical and clinical development could, within the next 5 or 10 years, become locked into a trajectory toward negative outcomes that will be hard to alter.

Based on the outside view estimates above, large-scale impacts of a new neurotechnology would not occur for at least 10 years after its initial demonstration in humans, and typically more like 25 years. No foreseeable neurotechnology is agentic or self-replicating, so neurotechnology is unlikely to directly cause rapidly escalating catastrophes analogous to AI misbehavior or engineered pandemics. Unless an extrinsic shock like global nuclear war slows or stops neurotechnology development, most effects of neurotechnology will occur at the pace of technology adoption.

However, it's harder to influence the use of a technology the more widely adopted it is. Establishing governance early is wiser than trying to convince frontrunners to slow down once they have established a lead. And as discussed more in the Tractability section below, intellectual property rights afford significant control over the pace and direction of neurotechnology development. (For reference, U.S. patents generally last 20 years, and trade secrets can be protected indefinitely.)

Thus intervention in the development of a neurotechnology around the time of its first demonstration in humans may be advantageous to guide its development trajectory towards good outcomes. Doing so later, after a technology developer has gained significant market power or a neurotechnology has become DIY-able, may be impossible.

Are there any neurotechnologies that may be at such a point in their development?

The most well-publicized examples are high-bandwidth sensory and motor cortical BCIs. Benefiting from over 30,000 hours of clinical data establishing safety and efficacy of cortical

arrays for computer interaction, they have garnered in the 100s of $M in commercial investment in the past 5 years. Their development may be further accelerated by minimally invasive approaches like stentrodes. Consumer desirability is unknown at this point, and all companies in the area appear to be following the typically-decade-long medical device approval process rather than going direct-to-consumer. While most outcomes from adoption of cortical BCIs are likely to be extremely positive, especially for disabled populations in the nearer-term, their interplay with internet communications raises questions about potential value shift, and their security implications regarding AI are unknown.

Ultrasound neuromodulation is another example. While no clinical uses for it have yet been established, its clearly-perceptible effects may elicit a large market. And the user-steerability of transcranial (i.e. noninvasive) ultrasound systems and their DIYability may enable rapid iteration and development. Regulatory intervention could easily attenuate its uptake, however, as could the relative dearth of safety data. While it shows great promise for treating neurological and neuropsychiatric disorders, its potential to noninvasively alter mood, affect, and other contributors to subjective well being raises questions about unintended value shift, wireheading, and abuse by governments.

Biologic neurotechnologies like monoclonal antibodies, gene therapies, and cell therapies (and arguably biosynthesized small molecules) are still early in preclinical development, but their potential effects are extremely broad, and more than other neurotechnologies they have a rapid potential path from scientific publication directly to DIY use. Neurotechnologies like cortical BCIs rely on generally inaccessible technologies like microfabrication and neurosurgery. But mail-order biotechnology equipment and reagents may be sufficient for an individual to replicate biologic neurotechnologies. This could facilitate rapid development and adoption. Much of this might be positive, such as being able to reproduce the effects of seemingly-beneficial mutations like FAAH-OUT, whose carriers reportedly feel pain sensations but don't experience suffering from them. But rapid development and adoption also risks unintended value shift.

In addition to these specific examples, there is the possibility of serendipitous discovery or stealthy development of a neurotechnology with potential for massive impact. Several companies known to the authors under NDA are developing neurotechnologies not listed in the section above.

Differential development timelines

The previous section was concerned with the impact neurotechnology might have if it develops along its current trajectory. But to what degree could concerted effort alter this trajectory and differentially accelerate the creation of neurotechnologies relevant to pressing concerns like AI safety?

Neuroimaging technologies could potentially be useful for AI safety even with only a small number of users. A takeoff scenario sensitive to initial conditions, like imitation learning on a single human's values, could be radically different in the presence or absence of a

neuroimaging system affording better access to the single human's moral judgments. And the longer AI development timelines are, the more influence neurotechnology may have.

Given that (1) widespread adoption is not necessary for a neuroimaging technology to contribute to AI safety and (2) the pace of neurotechnology development could be accelerated with the right interventions (see the Tractability section below), it is not unreasonable to estimate that with concerted effort, neurotechnologies currently in preclinical development could be developed within 10 to 20 years to the point that they would meaningfully benefit AI safety. Ultrasound and photoacoustic neuroimaging technologies are strong candidates for such a technology, which could potentially deliver 10x improvements in spatial and temporal resolution over fMRI in a portable device.

Neglectedness

While ~$20B/year goes toward funding neuroscience overall, only around ~$4B/year goes toward neurotechnology development, and almost no resources go toward differential neurotechnology development or governance.

Neuroscience (not neurotechnology) research landscape

Estimates of global government funding for neuroscience are not readily available. The major funder in the U.S. is the National Institutes of Health (NIH). The NIH provides funding for neuroscience research in the range of ~$5B to ~$10B per year. The US's Defense Advanced Research Projects Agency (DARPA) also funds neuroscience projects in the $100M per year range. The European Research Council provides funding for neuroscience in the €100M per year range. The Human Brain Project in Europe committed ~€1B to neuroscience in 2013, though it is regarded as not having produced valuable outcomes. China's funding landscape is more opaque. This report from CSET is the most detailed analysis we are aware of. It suggests China is spending in the 100s of $M per year on neuroscience research, with infrastructure outlays in the billions of USD to establish research centers in some cases.

Some non-governmental organizations alsof fund neuroscience research. Exemplars include The Allen Institute for Brain Science, which spends about $100M per year (some of which is from the from the NIH); two Max Planck Institutes, which as a rough estimate may spend ~$100M per year between them; the Howard Hughes Medical Institute's Janelia Research Campus, which has a ~$130M annual operating budget; the Kavli foundation, which has endowed various neuroscience institutes in the tens of $M range; and a number of disease-specific groups like the Michael J Fox Foundation, which has funded of $1B in Parkinson's research since 2000.

Altogether a reasonable **lower-bound estimate of global governmental and non-profit funding for neuroscience (not neurotechnology) is $20B/year** in the past five years. This

figure does not include research into neuroscience-enabling technologies like machine learning algorithms or laser miniaturization.

Neurotechnology research landscape

While all neuroscience research is relevant to the development of neurotechnology to some extent, most effort in neuroscience is not directly focused on developing neurotechnologies.

In terms of government funding for neurotechnology specifically, the US's BRAIN initiative, started in 2013, is an instructive example. Motivated in part by the lack of concerted development of neurotechnologies (as opposed to basic neuroscience), the BRAIN initiative disburses funding in the $400M/year range, increasing over time, having given ~$2.4 billion to date. The BRAIN initiative is itself funded by the NIH, National Science Foundation, DARPA, and several nongovernmental organizations.

While the BRAIN initiative funds some basic neuroscience research, it is mostly focused on research relevant to neurotechnology development, so we may consider the BRAIN initiative's budget as a lower bound of the amount of U.S. government funding devoted specifically to neurotechnology.

China's neuroscience funding is reportedly more focused on neurotechnology than basic research, especially BCIs, including investments in key research infrastructure like nonhuman primates. But there remains much uncertainty about how China's neurotechnology funding is being directed and the effectiveness of the R&D it is funding. An upper-bound estimate of Chinese government funding for neurotechnology-specific R&D, including infrastructure investment, is $500M/year.

Of the nonprofit funders listed above, it is difficult to estimate how much is devoted to translational research relevant to neurotechnology development. The Allen Institute has historically focused on basic neuroscience research, while Janelia has focused on neurotechnology tool development throughout its history, e.g. the Neuropixels project and GECI development. Disease-specific donors often fund projects characterizing diseases or doing other basic research rather than building neurotechnologies. A generous estimate would be that ~$300M (roughly double the Janelia budget) per year of nongovernmental funding goes toward neurotechnology development.

Assuming the U.S. government, Chinese government, and high-profile nonprofits make up the majority of neurotechnology funding, altogether this suggests an estimate of ~$1.5B/year of global governmental and non-profit funding research funding is focused on neurotechnology development. This is around 7.5% of the total spent on neuroscience. That figure broadly accords with the experience of the authors (based on publications, conference posters, presentations, and grants observed on various neuroscience topics) that most neuroscience effort goes toward using existing tools to explore the brain rather than building new tools that advance neurotechnology.

The majority of investment into neurotechnology development comes from for-profit enterprises. However, the vast majority of this investment is focused on drugs for neurological and neuropsychiatric disorders. Funding for neuro drug development is on the order of $10s of billions per year, (source, fig 1a) though much of this is conditional on achieving milestones and will never actually be disbursed. (Neuro drugs historically have about a 15% success rate.) Around 200 drugs are currently in development for various mental health disorders and around 500 for neurological disorders. Investment into non-drug neurotechnologies is significantly less. Of the ~$3.4B in neurotechnology investment announced in Q4 2021, only around ~$850M (~25%) was to companies focused on non-drug, central nervous system neurotechnology (source, page 12).

In sum, **a rough estimate of the amount of global investment from all sources in non-drug, central nervous system neurotechnology development is ~$4B/year**.

Efforts toward differential neurotechnology development or governance

Differential neurotechnology development is a slippery phrase, but herein we take it to mean efforts whose *stated* goal is to ensure beneficial neurotechnologies are developed before, and ideally in lieu of, harmful ones.

By this definition, arguably the most prominent effort in differential neurotechnology development is Neuralink. Explicitly motivated by AI safety concerns, Neuralink's mission is to allow the "merging" of humans with AI. What precisely this means has been a subject of debate. Neuralink has received $363M in investment since its founding six years ago.

Open Philanthropy funded Prof. Ed Boyden's lab, which focused on neurotechnology, in 2016 and 2018 for around $6M total, but it does not seem to have been motivated by differential neurotechnology development concerns. (source)

In terms of governance, in the U.S. the Food and Drug Administration (FDA) has jurisdiction over neurotechnologies that it considers to be for medical purposes, and controls the availability of such neurotechnologies for the purposes of improving public health. The Drug Enforcement Administration (DEA) in the U.S. and similar agencies around the world have mandates to reduce the availability of chemical neurotechnologies deemed public health risks. The net benefits of these regulators' actions is subject to much debate.

There seems to have been little proactive legislation or regulation around future consumer, non-drug neurotechnologies. The only concrete example we are aware of is Chile adding "neurorights" to their constitution in 2021. However, regulators like the U.S. Consumer Product Safety Commission could likely bring such new technologies under their jurisdiction after their development.

The field of academic neuroethics is arguably a proto-governance effort, though its influence is unclear, as is how much funding it receives. Professional associations like the IEEE also have efforts in neuroethics.

In total, the efforts described above are not well-measured in dollar terms, but they are few and leave much for a new philanthropist to do.

Tractability

While transformative neurotechnology will take decades to arrive at the earliest, there are opportunities today to proactively steer the field.

Fund research (unfinished)

Research into the cost-effectiveness of neurotechnology development for global health and wellbeing, which we did not attempt above, would resolve significant uncertainty about its value.

Run a proper survey of neurotechnologists on timelines. Examples of specific questions that deserve careful forecasting include:

- When will the Information Transfer Rate (or ideally a better metric) from a BCI exceed what's achievable by typing and speech?
- When will a para/tetraplegic human exceed a world track record using BCI-controlled prostheses?
- When will a neuroimaging system be able to preemptively predict a user's moral judgements (in a binary prediction task) with >90% accuracy?
- What is the probability that by 2050 there exists a frontline treatment for anhedonia with >90% success rate?
- When will the first non-drug, consumer (i.e. non-medical device) neurotechnology reach 1M users?

More research like this: https://pubmed.ncbi.nlm.nih.gov/30455187/

CSET's STI initiative

Fund or sponsor neurotechnology research that's relevant to AI safety. This topic is entirely off the radar of most neurotechnology researchers. **Funding consciousness researchers, AI safety researchers, and neurotechnologists to jointly develop concrete experimental plans, and then funding those experiments, would be a valuable addition to humanity's AI safety portfolio.**

Perform advocacy

Engagement with the FDA on their categorization of devices for general wellness or with the DOJ on their regulation of future neurotechnologies with abuse potential (i.e. avoiding another drug war) may be fruitful. Industry groups like IEEE Neurotechnologies for Brain Interfacing Group, which may have significant influence over industry standards in the future, may be effectively influenceable given how few stakeholders are engaged with them at present. Surgical boards are another potential advocacy target.

The first step for a new philanthropist toward any of these would be to fund research into policy and advocacy levers on neurotechnology.

Build infrastructure

The neurotechnology field is bottlenecked by poor infrastructure in multiple places. Building and controlling key infrastructure could allow a new philanthropist to facilitate beneficial neurotechnology research and hinder risky research.

One type of infrastructure is open-source software. Releasing top-quality open-source software for use in neurotechnology products can preclude the development of closed-source software and keep important aspects of neurotechnologies transparent. Examples might include open-source BCI operating systems or simulation packages for estimating safety of new stimulation patterns. EA-aligned software organizations like AE studio are equipped to do this and have fundable projects in this vein. The degree to which this infrastructure is valuable will depend on the specific neurotechnology and threat model.

Another piece of critical “infrastructure” is clinical cohorts. The pace of clinical trials and trial recruitment is a large factor of the overall pace of neurotechnology development. Providing clinical subject recruitment services to beneficial projects is another means of differentially steering the progress of neurotechnology. This is especially true for neurotechnologies that are not targeting diseases and will require clinical cohorts of healthy subjects. This is not hugely dissimilar to the 1Day Sooner challenge trial project that was funded by Open Philanthropy among others. A concerted effort to gather neural tissue samples into a biobank would also facilitate valuable research into neurological disease, and we are aware of interested parties for pursuing such an effort.

Control key IP

Unlike in the field of AI algorithm development, control over key intellectual property (IP) in neurotechnology affords private actors, including a new philanthropist, significant influence over the use of that technology.

IP infringement in electronic devices and biotechnologies, the key components of most neurotechnologies, is easier to ascertain and litigate than in software. Historically the power of

IP rights in neurotechnology has enabled anticompetitive behavior, with e.g. large companies acquiring and shelving patents from smaller competitors. Strategic acquisition of key IP (or controlling equity in its owners) by a holding company or nonprofit entity, followed by judicious licensing and monitoring of the use of that IP, could be a viable strategy for a private philanthropist to establish a “veto” over dangerous developments in neurotechnology. Market incentives will not necessarily pull the development of neurotechnology in positive directions, and establishing this veto power early may be prudent. Much more analysis would be required to determine the value and tractability of such a strategy.

TODO add benefits to ameliorating the risk that Neuralink or another large player becomes a bad actor in the space.

Stewardship of key IP can also differentially facilitate positive developments in neurotechnology. This is similar to software patent reform that Open Philanthropy has looked into in the past. A private philanthropist could, for example, fund a patent pool. Starting one would likely require total costs of <$100k. Starting a patent pool becomes challenging if any one company in a market has outsized market power, so early action may be sensible. Or a new philanthropist could acquire IP from dying startups to prevent it from being acquired by patent trolls and slowing progress in the field. TODO $100ks to buy some useful IP (or get an equivalently useful license) for philanthropists.

Differentially develop beneficial neurotechnology (unfinished)

The best way to ensure differential development of positive neurotechnology development is to build it yourself.

Neurotechnology development is pulled by funding, and funding for significant technology advances is pulled mainly by reimbursable medical indications.

TODO add FDF idea and FROs

From Quintin:

> A thought I've had about differential technical development in neurotechnology. There are known examples of especially suffering-related advances where dual use capability does not appear to be inherent to the advance.
>
> For example, if you generate therapeutics which target concurrent FAAH-OUT microdeletion + FAAH SNP, that doesn't disclose a way to cause extremely high pain sensitivity, at least by NOT affecting FAAH or other obvious genetic intervention.

> One could argue that in understanding its downstream implications for anandamide that does create a certain info hazard, but certainly not to the same extent that direct research on the mechanism first would.
>
> Why this is important is it biases me towards putting resources on the margin into interventions which have been discovered rather than engineered. I think there are enough low hanging fruit out there to fund in these categories (cluster headache treatments, novel chronic pain drugs, etc.) that a certain safety-aware donor pool could focus on.
>
> Another comment on differential development: when it comes to concerns about wireheading, it seems clear that some pleasures are 'healthier' than others. Contrast whatever the neural correlates of a runner's high are with those of fentanyl. Stimulating the release of endogenous endorphins seems clearly better for the organism than exogenous drugs, so that biases one approach to neurotech development over others (e.g. biophyiscally-safe stimulation over drugs, in this domain at least).

Acknowledgements

Special thanks to

- Parth Ahya
- Mike McCormick
- Ozzie Gooen
- Robert Long
- Mackenzie Dion
- Eliana Lorch
- Trenton Bricken
- Vishal Maini
- Steve Byrnes
- Quintin Frerichs
- Evan Miyazono

for insightful comments and useful suggestions.

Notes (won't be published)

A New X-Risk Factor: Brain-Computer Interfaces (LW, 10th Aug 2020)

- "This paper will outline how the development and widespread deployment of BCIs could **significantly raise the likelihood of longterm global totalitarianism**. We suggest two

main methods of impact. **Firstly, BCIs will allow for an unparalleled expansion of surveillance**, as they will enable states (or other actors) to surveil even the mental contents of their subjects. Secondly, BCIs will make it easier than ever for totalitarian dictatorships to police dissent by **using brain stimulation to punish dissenting thoughts**, or even make certain kinds of dissenting thought a physical impossibility."

BCI and AI Alignment (28th Aug 2021)

- "Whole-brain emulation (henceforth WBE) (with the emulations being faster or cheaper to run than physical humans) would likely be useful for AI alignment if used differentially for alignment over capabilities research"
- "no clear qualitative change in the way humans interact with AI systems, and create no differential speedup between alignment and capabilities work."

Using Brain-Computer Interfaces to get more data for AI alignment (6th Nov 2021)

- An unpublished paper from Borg and Sandberg breaks down the relevance of neurotechnology to AI safety into 3 categories:
    - Enhancement
    - Merge
    - Learn human values
- "A key component of various AI alignment proposals is teaching AIs something about humans: how humans think, or why we do the things we do, or what we value. AIs have a limited amount of data from which to learn these things. BCI technology might improve the quantity and quality of data available for learning."
- 3 possible ways of using such data
    - Learn models that imitate or predict neural signals
        - Understand the proposal, don't get why useful. Read Hubinger's paper.
        - This is broadly similar to Gwern's proposal
    - Get a richer signal to help models improve in debate or approval-based amplification
    - Learn the human value function

This twitter thread with Rob Bensinger of MIRI

- Reasons Rob isn't excited about BCI:
    - Brain-computer interfaces are very unlikely to make humans competitive with AGI.
    - Concerns about unintentional value shift
    - Accelerate capabilities research
        - Counterpoint: so does literally every good thing. The argument you have to make is whether it *differentially* accelerates capabilities research.
            - E.g. maybe neurotech enables neuroscience that gives researchers ideas for more powerful AI algorithms, but doesn't give them countervailing insight into human value systems

Boston, Superintelligence, Chapter 2

- Just says BCI isn't a good path to superintelligence

https://intelligence.org/ie-faq/#WhatIsGreaterThan
- "[BCI] might hasten the arrival of an intelligence explosion, if only by improving human intelligence so that the hard problems of AI can be solved more rapidly."

Steve Byrnes' Intro to Brain-Like-AGI Safety:
https://www.alignmentforum.org/s/HzcM2dkCq7fwXBej8
- Mentions differential tech development issue of figuring out human motivation systems before figuring out how general cortical learning algorithms work
    - Race dynamic?
- Related: Anthropomorphic reasoning about neuromorphic AGI safety cites EY and NB as saying neuromorphic AGI is the most risky, but the authors disagree.

From Yudkowsky:
    - "outer optimization even on a very exact, very simple loss function doesn't produce inner optimization in that direction"
    - 2 paths to alignment:
        - Create the perfect AI God, in one try
        - Create corrigible AI
            - This includes stopping before superintelligence as a trivial solution
    - A lot of what he says is "we can't access the value function". If we could, it could revolutionize the outer alignment problem, though wouldn't obviously help the inner alignment problem

CSET's China AI/BCI report
- Recommend creating a scientific and technical intelligence organization