

Enhancing human intelligence with neurofeedback

Anthony Kirilov
[Lys Technologies](#)

Nicole Dobрева
[Lys Technologies](#)

Thomas Rialan
[Lys Technologies](#)

We are developing neurofeedback protocols for non-invasive Brain-Computer Interfaces (BCIs). These protocols aim to enhance user productivity and effective intelligence. Preferentially providing this product to AI safety researchers will accelerate their research program, thus addressing the concerning gap between AI capabilities, and AI alignment progress. Commercial viability of this product is likely in the long run because consumers and workers in many professions with high marginal returns to focus and intelligence would benefit from using it.

1. Problem overview

As of January 12th 2025, prediction markets expect artificial general intelligence (AGI) by 2031, with a 25% chance by mid-2027 [1]. There is extensive literature making the case that developing AGI has worrying consequences [2,3]. The purpose of AI alignment is to help us get to AGI in such a way that avoids these risks. However, it seems like the rate of progress in AI capabilities substantially outstrips that of alignment, suggesting that we are not on track to achieve this goal [4,5].

It might be possible to speed up progress in AI alignment if someone invented a device that substantially increased the effectiveness of AI alignment researchers; perhaps by enhancing their concentration, or providing live feedback on their research thought process [6].

2. Our solution

High-level description

Fast and accurate feedback loops are very helpful for quickly solving a problem. If a researcher could wear a BCI that would track his or her thought processes and send text messages with feedback signals such as “You lied to yourself regarding point P” or “You just violated the planning fallacy” this would increase their effectiveness [7,8].

Such a feedback loop could be built in practice if we could develop: (a) a system for measuring and interpreting brain waves with sufficient accuracy, [9] and (b) a collection of protocols which take said brain waves as input, and output a feedback signal to the user. This is known as neurofeedback, a kind of biofeedback, which teaches self-control of brain functions to subjects [10].

Example neurofeedback protocols

“The first principle is that you must not fool yourself – and you are the easiest person to fool.”
– Richard Feynman, *Cargo Cult Science*, Caltech commencement address 1974. [11]

There are other various neurofeedback protocols that target increased activation in areas of the brain associated with working memory, attention, and episodic memory. Randomized control trials have shown that attention and episodic memory can be improved moderately (15-30%)

by using fNIRS neurofeedback protocols [12]. For problem solving, a more prescient issue might be making unjustified assumptions without realizing it [13]. We take it as a given that alignment researchers don't want to fall for this. If a user were encouraged to state her assumptions (either verbally or mentally), it's conceivable that it would be detectable when she was lying to herself. Real time feedback can be provided via an application on AR glasses, smart watch, or other devices.

Here is roughly how one might design this protocol:

1. Training Phase: Subjects engage in controlled problem-solving tasks where we can verify the presence of self-deception.
2. Pattern Recognition: Our neural networks learn to identify characteristic patterns in prefrontal cortex activation associated with self-deception.
3. Validation Phase: Testing against known cases of reasoning errors.

It is hard to estimate how much neurofeedback protocols would increase a researcher's productivity. The effects of boosting attention and memory might not be so dramatic - if you are twice as focused, you produce twice as much (or likely some smaller multiple). However, the benefit of 'not fooling yourself' is much larger as it could save you months of research down the wrong path.

Scientific foundations of our solution

When a part of your brain is active, the oxygenation level of the blood in that region of the brain is higher than elsewhere in the brain. This change in oxygenation can be detected, as near-infrared (NIR) light is not absorbed in the same way by oxygenated and not-oxygenated blood (oxy and deoxyhemoglobin). So, when we place sources of NIR light on your head and detectors that count the photons that come back, information about brain activity is encoded in these photon counts. It is already possible to purchase functional Near-Infrared Spectroscopy (fNIRS) headsets that measure exactly this. Relating photon counts to changes in oxygenation is known as solving the diffuse optical tomography (DOT) inverse problem. [14] It takes the following form:

$$Y = AX + e$$

where Y is the fNIRS data, X is the vector of oxygenation changes on the cortex, e is an error term (introduced due to linearisation) and A is known as the sensitivity matrix. See [11] for a full discussion of this.

Setting up this inverse problem, i.e. computing the matrix A , requires 3D computer simulations of shining NIR light through a realistic model of the head (including the skin, skull, white matter, gray matter, and a few other layers). Solving this problem, i.e. finding X from Y and A , requires inverting the matrix A . This problem is ill-defined due to the dimensionality of the problem but can be approximated by various methods such as minimum-norm estimates or Moore-Penrose inverses. Once you've solved the inverse problem for X , this is high-quality training data for a neural network to, e.g. perform the classification task "was the user lying to himself". A convenient side-effect of shining NIR light on your brain is that this stimulates the regions that absorb the light and there is evidence this enhances cognition.

Our solution consists in setting up this pipeline: collect fNIRS data, run it through a software pipeline that solves the DOT inverse problem with unprecedented accuracy, use this data to train neural networks that will categorise patterns of thought, and provide feedback to the user

on their thought pattern. The user would be wearing a cap to which the fNIRS light sources and detectors will be attached.

Competitive advantages and IP potential

Corporate structure benefits: Many approaches to AI safety, by their very nature, tend to be set up as non-profits. Because our proposed product will eventually provide value to a large segment of the consumer market, we've set up Lys Technologies as a for-profit corporation and will be able to benefit from venture capital investment.

Talent: Because developing a BCI requires a different set of skills to developing AI systems, we don't compete for talent with the large labs, which will help keep costs low compared to, e.g. starting a mechanistic interpretability or safety evals business. It's also less likely to hurt alignment research at the margin than it would be to reallocate people from an existing and perhaps promising line of research.

Technical secrets and IP potential: We've already developed a novel approach to brain signal processing that allows us to achieve high-precision categorisation of brain signals [15]. Because the topic of human intelligence has historically been rather taboo, there is good reason to believe that this is a field ripe with secrets, i.e. IP potential [16]. Also, unlike in AI, neurotechnology patents are enforced and enforceable.

3. Pilot experiment or demo

Our goal for the hackathon was to think through how we can apply our work to the question of AI safety.

Until a few days ago (written Jan. 13th 2025), we had mostly been working on this because we were personally excited by specific possibilities of BCIs and had not given much thought to the application of this technology to the question of AI safety. Our recent consideration of AI safety has led to, (a) the idea of preferentially providing our cognition enhancement tools to alignment researchers, and (b) doing some reading and thinking about AI safety that we hadn't all done to date.

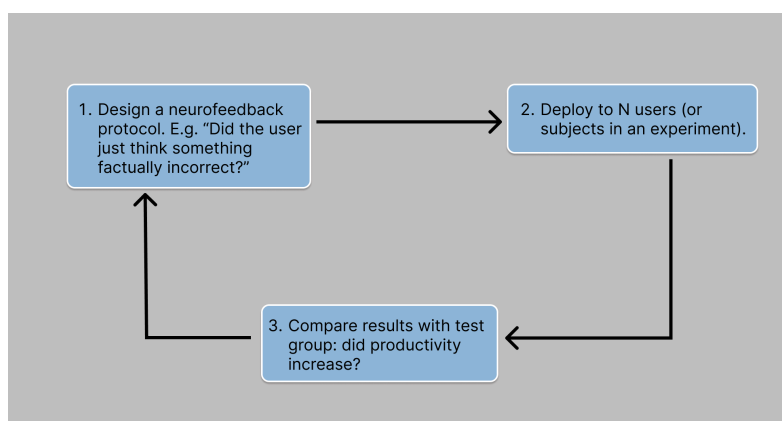


Figure 1 – Developing and deploying new neurofeedback protocols at scale could be viewed as analogous to A/B testing in software, but instead of optimizing the click-through rate, we optimize a measure of productivity.

Limitations and future development needs

There are still many engineering obstacles that need to be overcome: questions of latency and real-time classification (as opposed to approaches that only work in the lab), experimenting with hardware to detect lower amplitude changes in oxygenation, and a number of signal processing steps, including but not limited to the algorithms that solve the DOT inverse problem. [17,18]

4. Process

Timeframe	What we will do
Next 3 months	Buy an fNIRS headset, collect and analyse data from one neurofeedback protocol.
2025	Experimentally establish a suite of effective concentration and rationality-enhancing neurofeedback protocols. Work on the brain wave data pipelines.
2026	Work on the form factor for the device, sell the first few instances of the product. Meet prospective customers by attending relevant events: AI safety workshops, etc...
2027	Scale sales and production. Integration with AR glasses.

5. Impact on AI safety & key risks

Direct and indirect safety benefits:

Paul Erdős was a famously prolific Hungarian mathematician, comparable only to Euler for the volume of his production. He also took amphetamines for 25 years; this helped him maintain his intense productivity. When his mathematician friend Ronald Graham bet him \$500 that he couldn't stop taking drugs for a month, Erdős successfully quit for that time and then told Graham "You've set mathematics back by a month." Needless to say, his amphetamine consumption was resumed. What we need is to set alignment research forward by a decade. These are the direct safety benefits we are striving for.

There may be a few indirect benefits from our approach. The first is that because using specific neural circuits in your brain reinforces them, we might expect that when a user has worn the BCI for sufficient amounts of time, they will internalise the neurofeedback protocols and no longer need to even wear the device. On the more emotional and social side, the excitement that will stem from having an early-access device may have positive spillover benefits [19]. Thirdly, one traditional avenue for AI alignment is collecting large amounts of human judgment data and using that as feedback for the model (e.g. RLHF). It may be possible to collect large datasets of human moral judgments in a way that could be used to steer AI systems.

Potential risks and mitigation strategies:

The risks with new technologies are often many, and neurotechnology is no exception. Most straightforwardly, if this can help AI alignment researchers, then if it gets into the hands of AI capabilities researchers it will help *them*. To mitigate this we'd do a phased rollout with vetted research institutions that have a strong track record of AI alignment work. Contractually enforceable terms and conditions could restrict the conditions on use. Another risk is that by learning more about the brain with our research, we might discover insights into building more powerful AI systems, as neuroscience has in the past. This particular risk can be mitigated by the careful protection of IP. Thirdly, there are serious privacy concerns with a device that may eventually be able to read the thoughts of its users. The associated risk of "mind control/mind policing" scenarios needs to be carefully navigated. The key mitigation approach here will be to aim to perform all of the data processing client-side, without needing to communicate with a server.

Ultimately the number of unknowns surrounding this tech makes it hard to accurately weigh the risks and rewards without further research. [20]

Present success metrics and KPIs:

We've worked on this problem for the last six months. At present our research is progressing well and we mostly see a clear path forward for building what we've described. The main constraint is time, *i.e.* funding. We currently have enough money to keep working on this for ~6 months; having an 18 month runway would give us ample time to make serious progress.

Appendix

- [1] Metaculus market for "When will the first general AI system be devised, tested, and publicly announced?"
- [2] *Superintelligence: Paths, Dangers, Strategies*, Nick Bostrom, 2014.
- [3] The Sequences
- [4] [*The Field of AI Alignment: A Postmortem and What To Do About It*](#), johnswentworth, 2024.
- [5] [*Against Almost Every Theory of Impact of Interpretability*](#), Charbel-Raphaël, 2023.
- [6] It might be argued that if something can enhance the effectiveness of AI alignment researchers, this would also enhance the effectiveness of AI capabilities researchers. This is discussed in section 5.
- [7] Accurate BCI based lie detectors already exist, to our knowledge no-one has made a lie detector that is aimed at detecting when one lies *to oneself*.
- [8] [*Rationality is Systematized Winning*](#), Eliezer Yudkowsky, 2009.
- [9] There have been many approaches to this over the years with devices such as EEG, MEG, fMRI, Utah arrays, stentrodes and fNIRS.
- [10] Definition from Marzbani H, Marateb HR, Mansourian M. [*Neurofeedback: A Comprehensive Review on System Design, Methodology and Clinical Applications*](#). Basic Clin Neurosci. 2016
- [11] It is nice when quoting this to self-reflect, "am I fooling myself *right now*". The biggest question here isn't so much do-ability in principle, but timing. Perhaps this can only be done on the hardware we'll have in 15 years? We think not.

- [12] [Cognitive Training with Neurofeedback Using NIRS Improved Cognitive Functions in Young Adults: Evidence from a Randomized Controlled Trial](#), Nouchi et al, Brain Sci., 2022.
- [13] Making unjustified assumptions, jumping to conclusions, etc.
- [14] We refer the reader to [Cai, Z., Machado, A., Chowdhury, R.A. et al. Diffuse optical reconstructions of functional near infrared spectroscopy data using maximum entropy on the mean](#), Nature, 2022, for a first look at the extensive literature on this subject.
- [15] We chose not to disclose the details of this signal processing for competitive reasons.
- [16] Chapter 8, *Zero to One: Notes on Startups, or How to Build the Future*, Peter Thiel and Blake Masters, 2014.
- [17] Our current research is about demonstrating the effectiveness of our approach on data collected in a laboratory setting. Deploying systems to users will require addressing questions of power consumption and latency of the signal processing. We have ideas for how to address these but haven't yet gotten around to implementing them.
- [18] Hardware needs to be detecting the "initial dip" in the blood-oxygen-level dependent image. See Zafar A, Hong KS. [Detection and classification of three-class initial dips from prefrontal cortex](#). Biomed Opt Express. 2016.
- [19] Generating positive emotions in the alignment community seems like a valuable side effect.
- [20] Many ideas from this section have been [discussed in the broader context of neurotechnology by M.W. Cytkovic on the EA forum](#).