The Fate of Human Fallibility in the Era of Cyborgs:
Analyzing the State of Cognitive Biases in a Revolution of Brain-Machine Interfaces

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Austin Baney
Fall, 2019

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Austin B

Signature ____ Date 5/5/2021
Austin Baney

The Fate of Human Fallibility in the Era of Cyborgs: Analyzing the State of Cognitive Biases in a Revolution of Brain-Machine Interfaces

Amidst the Information Age, humanity is more closely connected with itself and the digital world than ever, with unprecedented access to information and ability to communicate instantly. However, if brain-machine interfaces evolve as predicted by neuroscientists and experts like Elon Musk, this connection could become dramatically deeper and more powerful than ever thought possible. If innovations proceed at this trajectory, human communication with computers and other individuals could be as quick, lossless, and effortless as simply thinking. Needless to say, widespread adoption of this technology would have a significant impact on humanity as we know it. How would a network of full brain-machine interfaces affect the way people are influenced by others and by their own cognitive errors and biases?

Brain-machine interfaces (BMIs) use arrays of electrodes to decipher, transmit, and induce neural activity, essentially reading information from, and even sending information to, the brain. These exist today, but with limited capabilities and applications. Most existing BMIs are used for medical treatment. For example, robotic limbs can record activity in the brain's motor cortex to be controlled at will. Cochlear implants record sound and stimulate the auditory nerve accordingly to restore hearing. There have even been successful experiments in direct brain-to-brain communication between rats[1], between humans[2], and even from human to rat[3]. However, current BMIs are only capable of transferring information with very limited fidelity. Elon Musk intends to evolve the state of brain-machine interface technology with his venture Neuralink. According to Musk, a great barrier to this innovation is the inability to record the activity of enough neurons. Modern scientists struggle to implement enough electrodes to detect and stimulate neural activity at a high bandwidth. However, Neuralink is beginning to push this boundary by building BMIs that use small multielectrode arrays to transmit across 3,072 channels, capable of recording and inducing neural activity at an unprecedented level of fidelity.[4] Nerualink's BMI technology has seen success with rats, but Musk hopes this is the first step towards something greater – a high-bandwidth, whole-brain interface, capable of transferring information directly between human brains and the computers.

Today, the rate and quality at which people communicate – with computers and each other – is bottlenecked by our physical limitations. Information flows quickly and with high-fidelity in our brains, but this flow is slowed and limited in complexity when we attempt to transfer the information to a computer or another person by means such as language. As Musk explains:

> *There are a bunch of concepts in your head that then your brain has to try to compress into this incredibly low data rate called speech or typing. That's what language is—your brain has executed a compression algorithm on thought, on concept transfer. And then it's got to listen as well, and decompress what's coming at it. And this is very lossy as well. So, then when you're doing the decompression on those, trying to understand, you're simultaneously trying to model the other person's mind state to understand where they're coming from, to recombine in your head what concepts they have in their head that they're trying to communicate to you. ... If you have two brain interfaces, you could actually do an uncompressed direct conceptual communication with another person.*[5]

With such a whole-brain interface, the only bottleneck to communication would be the rate and complexity of human cognition; people would, in essence, be able to share information as quickly and with as much detail as they can think. Anyone with a full BMI would be granted immensely improved access to information and ability to process it. People may become capable of sharing thoughts, emotions, memories, and experiences in full detail. Humans with a fully-realized BMI might even make such communications subconsciously, yielding an inconceivably deep integration with the digital world and other minds. To propose that such an innovation would transform society at large would be an understatement. Whether this transformation is more beneficial than dangerous is not so clear.

Such a powerful revolution in digital communication would certainly raise many ethical concerns. The security of a whole-brain interface would be paramount, and calls for careful consideration. A hacker who gains access to someone's BMI could also access and manipulate their thoughts in concerningly dangerous ways. Even if hacking is not an issue, what if a BMI user can authorize trusted individuals to access their interface? Predators and malicious entities could seek out suggestible targets, gain their trust, and abuse their access. Modern social media is already a conduit for much harassment and radicalization, but such dangers could multiply with global brain-to- brain communication. Furthermore, a sobering concern regarding a BMI network is how one's sense of identity could be affected. Could individuals deeply connected by the brain even call themselves individuals, or would an emergent shared identity prevail?[14] Questions like these should be thoroughly investigated to prepare society for this evolution of BMI technology.

Any large-scale technological revolution bears numerous ethical implications, but this particular advancement could even carry an impact on the flaws that have been part of humanity since the beginning. Humans have always been subject to cognitive errors and biases. Despite our advancements thus far, we seem likely as ever to fall prey to our misguided mental shortcuts, our tendency to conform to the norms of the group, and our desire to confirm our existing views. Could the advent of the whole-brain interface free us from these flaws and help people understand each other's points of view? Or would it become that much easier to reinforce our presumptuous beliefs and entrench into divisive echo chambers?

Consider the notion of mental models – a person's deep and unique cognitive representation of a concept. Different individuals tend to construct these models in vastly different ways, even when they are representing the same ideas. For example, Alexander Graham Bell based his mental model of the telephone on the human ear, while Thomas Edison took inspiration from the telegraph – two different approaches to conceptualizing the same technology.6 With the limited sophistication of modern communication, sharing mental models accurately is difficult. This creates barriers between individuals, and between groups with irresolvable shared mental models. Isolated communities may share a mental model that normalizes behaviors considered outside the community to be deviant or harmful. An example of such normalization is the practices of NASA engineers which led to the infamous Challenger shuttle disaster.[7]

Would problems stemming from incompatible mental models be solved by whole- brain interface technology? If any complex idea could be shared accurately – to computers or other people – could mental models could be sent directly between brains? Even a fully-realized BMI might not be capable of transferring deeply subconscious mental models, especially since long-term memories are often reconstructed when prompted with relevant stimuli, and change based on current assumptions and predispositions.[13] However, such an interface could perhaps aid in recalling memories with more accuracy by "jogging" the memory with proper stimuli, or even by recording and storing the memories in detail as they are formed. If this technology does indeed become capable of direct mental model transfer, it would open endless possibilities for deep collaboration and understanding between individuals and groups. In many industries, tacit knowledge and tricks of the trade requiring firsthand experience could suddenly be taught with ease. Ramez Naam, founder of Apex Technologies, pondered the effects of BMI-powered group thinking:

That type of communication would have a huge impact on the pace of innovation, as scientists and engineers could work more fluidly together. And it's just as likely to have a transformative effect on the public sphere, in the same way that email, blogs, and twitter have successively changed public discourse.[8]

On the other hand, given a deepened integration with the content we consume, escaping flawed mental models and normalized deviance could become more of a challenge than ever. Humans are already prone to subconsciously shift their views to align with the unspoken norms of the group. If this digital interface blends seamlessly with the mind, and complex mental models are shared among communities, people could be subject to drastic realignments of their views without noticing. A means of escaping patterns resembling normalized deviance is development of one's own "moral imagination" – the inclination to seek perspectives and consider possibilities from outside the dominant narrative.[6] While full BMIs could make it easier to diversify one's perspective in this way, suggestible individuals unaware of the flaws in the norms of their community may not even consider attempting this.

Even when acting independently of a group, human cognition is prone to adoption of flawed generalizations and assumptions. Every person, knowingly or not, uses heuristics – subconscious mental shortcuts taken in spite of uncertainty. This is often beneficial, helping us make quick decisions without getting bogged down in unnecessary complexity. However, a complacent lack of awareness regarding one's own assumptions leads to heuristic failures, which can cause even medical professionals to make harmful yet avoidable decisions.[9]

Would a future with worldwide neural interfacing see a decline in such heuristic failures? With a bolstered ability to find, process, and retain information, perhaps such crude mental shortcuts will become unnecessary. Even if humans still rely on them, heuristics could encode far more nuance with full BMIs than they do today. With more complex ideas readily accessible, we might stop making serious mistakes based on crude rules of thumb. One such flaw in human cognition is availability bias – the inclination to prioritize information that is familiar or comes easily to mind. BMIs could combat this bias by making a vast wealth of information available instantaneously.

A strategy for avoiding heuristic failures is known as "meta-cognition" – deliberate awareness of one's own decision-making and the uncertainties they tend to overlook.[9] With the adoption of whole-brain interfaces, would this practice become more easily adopted in humans? A boost in raw intelligence may not be enough; escaping the pitfalls of human error and bias requires a level of self-awareness and conscious attention to our own cognition. Even in an age of superhuman thought and communication, it could still be an individual's responsibility to keep oneself in check.

Perhaps one of the most powerful yet subtle influences on modern discourse lies with the widespread reliance on online platforms such as social media. Not only have we adapted to these platforms, but they have learned to adapt to us. Sites like Google, Facebook, and Twitter are known to filter the presentation of content to appeal to their users' preferences, not only contributing to the platforms' addictive nature, but also quietly steering their users toward views with which they already agree. Eli Pariser dubbed this phenomenon "filter bubbles" – in which online media entrenches its users in echo chambers populated with similar opinions.[10] In an interview with Quartz, Bill Gates says that social media "lets you go off with like-minded people, so you're not mixing and sharing and understanding other points of view ... It's super important. It's turned out to be more of a problem than I, or many others, would have expected."[11]

In an age of ubiquitous whole-brain interface technology, social media will be more present than ever, with the views of friends, public figures, and corporations constantly flowing through our very minds. If these streams of content become deeply ingrained in our human cognition, the insidious nature of filter bubbles could become concerningly powerful – especially if our BMIs let us browse this personalized content with little conscious effort.

Today, some critics assert that such algorithmically personalized content feeds are to blame for the problematic and divisive effects of filter bubbles. In anticipation of the neural interfacing revolution, could this problem be mitigated by enforcing a lack of bias in social media's content filtration algorithms? It may not be so simple, thanks to another flawed tendency in human cognition: confirmation bias. This is a person's subconscious inclination to search for evidence or opinions that confirm their existing beliefs, while at the same time, discrediting that which conflicts with their desire for self-affirmation. Confirmation bias is believed to be a strong underlying cause of the filter bubble phenomenon.[12] An enforced "fairness" of content filtration, then, may do little to quell this human desire to live in the comfort zone. Once the world of social media platforms becomes smoothly integrated into our minds, the influence of filter bubbles could become nearly inescapable. Although, as with most any pattern of bias in human cognition, the most effective tool to escape such influence is a conscious awareness of what drives our decisions and a deliberate effort to challenge and diversify our views.

The concept of complete digitized interfacing between human brains and computers sounds outlandish and frightening, but in a way, this innovation is nothing completely alien.

Some may be understandably hesitant to introduce a digital layer to their cognition, but this dependency, in essence, is already present. As observed by Elon Musk:

> *The thing that people, I think, don't appreciate right now is that they are already a cyborg. You're already a different creature than you would have been twenty years ago, or even ten years ago. You're already a different creature. You can see this when they do surveys of like, "how long do you want to be away from your phone?" and— particularly if you're a teenager or in your 20s—even a day hurts. If you leave your phone behind, it's like missing limb syndrome. I think people—they're already kind of merged with their phone and their laptop and their applications and everything.*[5]

Despite this analogy, the essence of our identity remains as "human" as ever, even in the Information Age. We do, however, increasingly depend on technology to interface with one another and collect information, and often take this for granted. As this dependency more deeply intertwines with human cognition, perhaps a mastery of moral imagination and meta-cognition will distinguish the humans from the "cyborgs".

A future with widespread access to fully-realized brain-machine interfaces opens countless possibilities for benefits and dangers alike. Today's human integration with the digital world already fosters innovation and understanding, as well as division and ignorance. After a BMI revolution, humanity will still interface with this same world, just more intimately and capably so. The stakes will be higher than ever as the tool of digital communication sharpens, but how we use this tool can still be our choice. Just as this breakthrough will make it easier for us to mislead one another and reinforce our biases, so too will it afford a greater capacity for community, progress, and awareness of our shortcomings.

Feedback

I would like to thank Michael Dellaripa for providing me with guidance throughout the process of establishing the scope of my topic, and pointing me towards useful information, especially regarding mental models and moral imagination. He spent time giving me advice for my draft as well as the final prospectus.

I thank Professor Michael Gorman for supplying insightful feedback for my draft. He helpfully advised me to consider the concept of reconstructive memory and its application to my topic, and also encouraged me to further explore ethical implications. I also thank Prof. Gorman for being an effective teacher of the STS concepts used as a framework for this paper.

Bibliography

1) Pais-Vieira, M., Lebedev, M., Kunicki, C., Wang, J., & Nicolelis, M. A. L. (2013, February 28). A Brain-to-Brain Interface for Real-Time Sharing of Sensorimotor Information. Retrieved from https://www.nature.com/articles/srep01319.

2) Grau, C., Ginhoux, R., Riera, A., Nguyen, T. L., Chauvat, H., Berg, M., ... Ruffini, G. (2014, August 19). Conscious Brain-to-Brain Communication in Humans Using Non-Invasive Technologies. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105225#s4.

3) Yoo, S.-S., Kim, H., Filandrianos, E., & Taghados, S. J. (2013, April 3). Non-Invasive Brain-to-Brain Interface (BBI): Establishing Functional Links between Two Brains. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0060410.

4) Musk, E., & Neuralink. (2019, January 1). An integrated brain-machine interface platform with thousands of channels. Retrieved from https://www.biorxiv.org/content/10.1101/703801v2.full.

5) Urban, T. (2019, November 6). Neuralink and the Brain's Magical Future (G-Rated Version). Retrieved from https://waitbutwhy.com/2017/04/neuralink- cleanversion.html.

6) Gorman, M. E., Mehalik, M. M., & Werhane, P. H. (n.d.). Ethical and environmental challenges to engineering. Englewood Cliffs, NJ.

7) Vaughan, D. (2016). The Challenger Launch Decision. Chicago, IL: The University of Chicago Press.

8) Naam, R. (2015, May 16). The Ultimate Interface: Your Brain. Retrieved from https://rameznaam.com/2015/05/15/the-ultimate-interface-your-brain/.

9) Cumbler, E., & Trosterman, A. (2018, September 14). The Psychology of Error. Retrieved from https://www.the-hospitalist.org/hospitalist/article/123482/psychology-error.

10) Pariser, E. (2012). The filter bubble: how the new personalized web is changing what we read and how we think. New York, NY: Penguin Books.

11) Delaney, K. J. (2017, February 22). Filter bubbles are a serious problem with news, says Bill Gates. Retrieved from https://qz.com/913114/bill-gates-says-filter-bubbles-are-a-serious-problem-with-news/.

12) Self, W. (2016, November 28). Forget fake news on Facebook – the real filter bubble is you. Retrieved from https://www.newstatesman.com/science-tech/social-media/2016/11/forget-fake-news-facebook-real-filter-bubble-you.

13) Roediger, H. L. (2002, November 2). Reconstructive Memory, Psychology of. Retrieved from https://www.sciencedirect.com/science/article/pii/B0080430767015217.

14) Trimper, J. B., Wolpe, P. R., & Rommelfanger, K. S. (2014, January 25). When "I" becomes "We": ethical implications of emerging brain-to-brain interfacing technologies. Retrieved from https://www.frontiersin.org/articles/10.3389/fneng.2014.00004/full.

Further Reading

- Therapeutic deep brain stimulation reduces cortical phase-amplitude coupling in Parkinson's disease
    - https://www.nature.com/articles/nn.3997
- Flipping the switch: Targeting depression's neural circuitry
    - http://emorymedicinemagazine.emory.edu/issues/2015/spring/features/brain-hacking/flipping-the-switch/index.html
- Deep brain stimulation offers hope for OCD patient
    - https://www.cnn.com/2014/06/24/health/brain-stimulation-ocd/
- Creating a False Memory in the Hippocampus
    - https://science.sciencemag.org/content/341/6144/387.long