NDSEG Fellowship Research Proposal

Identifying and Addressing Adversarial Stimuli in *C. elegans*

Ronak Ramachandran

November 3, 2023

Motivation One of the central aspirations of science has been to understand the human mind and replicate its strengths. Both neuroscience and computer science were founded with this purpose in mind, with neuroscience seeking to break the complexity of the brain into its simplest components [1] and computer science seeking to build complex "thinking" machines from simple parts [2]. The next frontier in this interdisciplinary mission will be bridging the gap between humans and computers. Already, companies like Neuralink and Synchron are exploring new human-computer interfaces for various medical applications. Some, however, intend to eventually use their brain implants to connect people more directly to the internet. The concept is exciting, but also concerning. It can be difficult to trust that companies like these will properly prioritize the safety of their products. Despite troubling outcomes in animal trials [3], Neuralink has been approved by the FDA for human testing [4], and their management has been rushing projects, encouraging unnecessary risks [5].

One risk that seems often overlooked is the issue of neurosecurity, the idea of protecting the neural machinery and autonomy of an internet-connected person with neural implants from malicious parties on the internet [6]. Recent encouraging results with medical neural implants lend to the belief that it might be possible to connect the brain to the internet without fully understanding how the brain works [7] [8]. However, it is easier to break a system than to fix it, and fixing a system usually requires knowledge of its inner workings. With this in mind, a necessary prerequisite to the safety of brain-internet interfaces is building better models of the human brain and a suite of neurosecurity innovations that can unobtrusively protect the brain from malicious input signals.

What might these malicious signals look like? In 1997, a Pokemon episode involving red and blue images flashing at twelve frames per second—a stimulus that would never arise in nature—caused at least 600 children in Japan to have seizures, convulsions, headaches, and vision problems [9]. The vast majority of these children had never experienced a seizure before, and the percentage of viewers affected was far higher than the prevalence of photosensitive epilepsy [10]. This particular incident was an accident, but one could imagine bad actors specifically designing more harmful stimuli. There are less dangerous and more correctable examples: optical illusions are an example of how neural heuristics can sometimes fail, resulting in unexpected outcomes. Humans aren't unique in this regard; drawing a line in front of a chicken can cause it to freeze in place for up to half an hour, an effect called tonic immobility [11].

Borrowing language from machine learning, in which adversarial noise can cause image classifiers to misclassify objects, I call these stimuli which result in maladaptive behavior *adversarial stimuli*. Here, maladaptive means anything directly in opposition with the survival and well-being of the individual being stimulated. Note that neural implants aren't necessary to provide someone with adversarial stimuli, but implants will make accidents like the 1997 Pokemon incident more likely and more dangerous. Millions of years of evolution have not prepared the brain for the new stimuli it will soon be expected to process. The goal of this research project is to demonstrate some risks and potential remediations of adversarial stimuli with the goal of advancing neurosecurity.

Proposal To make the project tractable, we will use the model organism *C. elegans* as a stand-in for humans and explore the foundational principles of neurosecurity that might apply to any nervous system. We chose *C. elegans* because the millimeter-long transparent worm has one of the best characterized and most easily manipulable nervous systems. Researchers have identified all 302 of its neurons along with the physical [12] and functional [13] connections between them, providing us the luxury of studying its nervous system at the level of individual synapes. The project consists of five stages:

Stage 1: Construct and optimize a predictive model of the nervous system of *C. elegans*.
Stage 2: Use that model to search for stimuli that result in maladaptive behavior in the model.
Stage 3: Test whether the maladaptive responses predicted by the model are seen in *C. elegans*.
Stage 4: Use the model to discover neural manipulations that block malicious control.
Stage 5: Test which of these remediations work in real *C. elegans* specimens.

With *C. elegans* as a testbed, the hope is that insights from this research program might help us design neurosecurity primitives that could be validated in more complex nervous systems (like rodents and primates) before being adapted to humans. This research program is ambitious and will likely take years to complete, but each stage will yield results that are useful in their own right and may spawn new research directions.

Ongoing Work With the guidance of Drs. Jon Pierce and Xuexin Wei, I've made progress on Stage 1, creating and optimizing a predictive model that simulates the *C. elegans* nervous system at neuron activation granularity. We want to understand how signals are propagated between and processed inside neurons, because we will need the model to inform the ways we modify the nervous system in Stage 4. Naively using deep learning to predict behavior directly from stimuli would be insufficient, for instance, because the black-box nature of the weights in artificial neural nets makes it unclear how modifications to your model correspond with modifications to the physical system you're modeling. For this reason, we use a regression-based model, though we may augment it with machine learning later. The goal of our model is to mimic the activation traces of every neuron in the *C. elegans* nervous system. Due to the difficulty of directly measuring neuron voltages, we use an extensive public collection of labeled whole-nervous system calcium traces [14] as a proxy for neural activation. We also have access to the graph of physical [12] and functional [13] connections between neurons, the connectome.

Imagine $f_i(t)$ is our predicted activation at time t for the neuron i. The first thing we might consider is the simple recurrence $f_i(t) = af_i(t-1)+b$, which has closed form $f_i(t) = b(1+\cdots+a^{t-1})+a^t f_i(0)$. This is a good starting point because it can model constant ($a=0$), linear ($a=1$), and exponential ($b=0$) behavior, all of which are exhibited by neurons at various points of time. In order to switch between these regimes, however, we would like a and b to be functions of the activations of neighboring neurons. For simplicity, we can imagine they are linear functions of the activation history of all relevant neurons going as far back as w_{rr} time steps for some w_{rr} we'll call the width of relevant retention. Finally, we can extend our recurrence to depend on multiple prior values of f_i, so $f_i(t) = a_0 + a_1 f_i(t-1)+\cdots+a_{w_{\text{sr}}} f_i(t-w_{\text{sr}})$ for some w_{sr} we'll call the width of self retention.

This results in the following model: Fix some neuron i. Let R_i be the set of neurons relevant to i (in our model, we consider i and all incoming/outgoing neighbors to be relevant to i) and let r_{jk} be the activation at time $t-k$ for the jth neuron relevant to i. Let f_{il} be the activation of i at time $t-l$. Then, we posit that f_{i0} is given by

$$f_{i0} = \sum_{j\in R_i}\sum_{k=1}^{w_{\text{rr}}}\left(r_{jk}c_{jk0}+\sum_{l=1}^{w_{\text{sr}}} r_{jk}c_{jkl}f_{il}\right),$$

for some constants c_{jkl}. Here, $r_{jk}c_{jkl}$ represents a_l. Our goal is then to learn the 3-dimensional tensor C given by c_{jkl}. If we pre-process our raw trace data, we can calculate the above formula for all t with just one matrix multiplication. Then, linear least squares regression explicitly solves for the C minimizing error. My implementation of this model here can be found here: `https://github.com/ironak/worms/blob/main/model.ipynb`.

This model has a number of advantages. (1) It's fast. The linear regression to calculate C takes less than half a millisecond per neuron, which will be helpful if and when the model might be applied to more complex nervous systems. (2) It's parallelizable. The C for each neuron can be found individually, and then the results can be combined to make a model of the whole nervous system. (3) Preliminary results show it fits the data surprisingly well. For instance, our model explains 76% of the variance in the activation of the neuron AVAR. Other neurons

fare similarly well, but AVAR is a standout example because it is well connected and well represented in the data we've been able to train on. We're still in the early stages of the project though, so further work is necessary to verify that the model is doing something non-trivial. I'm currently in the process of automating the search for the appropriate values of w_{rr} and w_{sr} for each neuron so that I can solve for C for each neuron and then use these C to extrapolate a small bit of raw trace data as far as possible. How long will the simulation stay close to the truth? Since the errors in the model will compound, a lot more optimization will be necessary.

How does this model compare with other models? One classic method for simulating time series data is the Auto-Regressive Integrated Moving Average (ARIMA) model. Applying ARIMA to individual neurons results in a very close fit for training data, but it doesn't generalize well: the predictions made by the model on test data are far from the ground truth, likely because ARIMA does not take into account signals from neighboring neurons. We're in the process of fitting an ARIMA model to multidimensional data, and we plan to also compare our model to others based on neural nets.

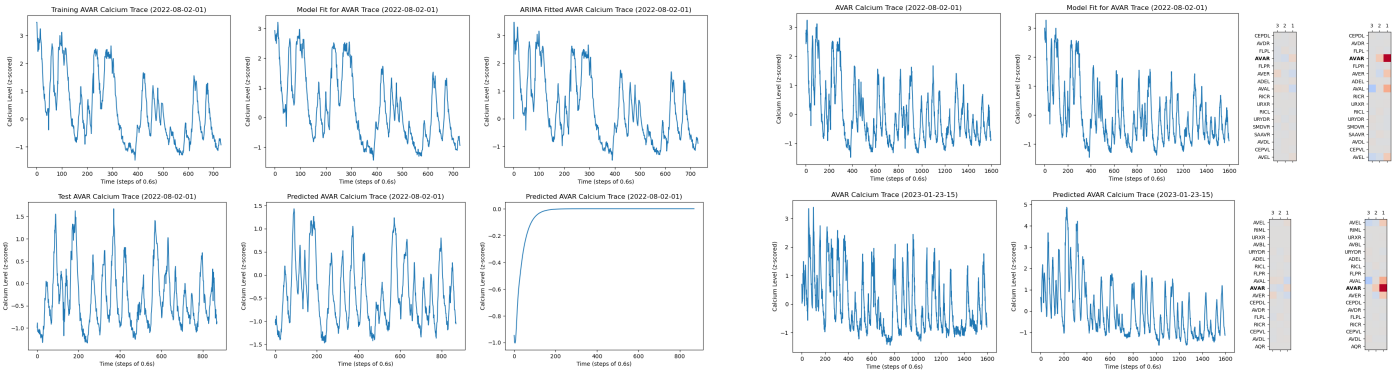

Figure 1: **Column 1:** Real calcium traces for the neuron AVAR. Top for training, bottom for testing. **Column 2:** Our model fitted to training data and its predictions for test data. Test prediction explained variance = 0.89. **Column 3:** ARIMA fitted to training data and predictions on test data. Test prediction is visibly worse, showing how crucial connectome data is. **Column 4:** Real AVAR calcium traces from two datasets. **Column 5:** Our model fitted to first dataset and predictions on second dataset. Explained variance = 0.76. **Column 6:** Model parameters visualized as 17x3x2 tensor with red for positive values and blue for negative. To apply the learned parameters to the second dataset, it was necessary to reorder entries and introduce one row of zero entries. Model parameters suggest AVAR has a strong non-trivial dependence on AVAL and AVER. All real trace data sourced from [14].

Potential Outcomes Ideally, our model lets us reverse-engineer stimuli that cause maladaptive behavior not previously observed. Perhaps a specific brief sequence of blue light pulses at light-sensitive neurons results in *C. elegans* moving in a circle indefinitely, forgoing food. Then, the fix may be as simple as habituation or might involve something more invasive, like ablation or silencing neurons through chemo- or optogenetics. With more adversarial stimuli, we may find our solutions relying on error correcting codes or cryptographic hardness assumptions for their safety guarantees. As we build up foundational principles and explore larger systems, Neurosecurity may become as prominent and complex as cybersecurity.

Next Steps Individual neurons are complex signal processing systems [15] which can modify their behavior through gene expression [16] and simultaneously handle multiple signals from various neurotransmitters [17]. Our model represents each neuron by a single number: its calcium activation. Future work might improve models by capturing this complexity. Additionally, unlike human neurons, most *C. elegans* neurons are non-spiking [18], so adversarial stimuli may look very different in humans. Even after this project is complete, a lot more work in neurosecurity will be necessary before the human brain will be ready to be connected to the internet.

References

[1] Hippocrates. *The Internet Classics Archive: On the Sacred Disease by Hippocrates*. Ed. by Francis (Translator) Adams. URL: https://classics.mit.edu/Hippocrates/sacred.html.

[2] A. M. TURING. "I.—Computing Machinery and intelligence". In: *Mind* LIX.236 (1950), pp. 433–460. DOI: 10.1093/mind/lix.236.433.

[3] Dhruv Mehrotra. *The gruesome story of how Neuralink's monkeys actually died*. Sept. 2023. URL: https://www.wired.com/story/elon-musk-pcrm-neuralink-monkey-deaths/.

[4] Sept. 2023. URL: https://www.reuters.com/technology/musks-neuralink-start-human-trials-brain-implant-2023-09-19/.

[5] Rachael Levy. *Exclusive: Musk's Neuralink faces federal probe, employee backlash over Animal tests*. Dec. 2022. URL: https://www.reuters.com/technology/musks-neuralink-faces-federal-probe-employee-backlash-over-animal-tests-2022-12-05/.

[6] Tamara Denning, Yoky Matsuoka, and Tadayoshi Kohno. "Neurosecurity: Security and privacy for Neural Devices". In: *Neurosurgical Focus* 27.1 (2009). DOI: 10.3171/2009.4.focus0985.

[7] Dan Hurley. "Brain-spine interface restores standing, walking, and stair-climbing after spinal cord injury". In: *Neurology Today* 23.13 (2023). DOI: 10.1097/01.nt.0000946564.14834.d4.

[8] Simanto Saha et al. "Progress in brain computer interface: Challenges and opportunities". In: *Frontiers in Systems Neuroscience* 15 (2021). DOI: 10.3389/fnsys.2021.578875.

[9] Benjamin Radford. "The Pokemon Panic of 1997". In: *Skeptical Inquirer* 25.3 (2001), pp. 26–31.

[10] G.F.A. Harding and P.F. Harding. "Photosensitive epilepsy and image safety". In: *Applied Ergonomics* 41.4 (2010), pp. 504–508. DOI: 10.1016/j.apergo.2008.08.005.

[11] Gordon G. Gallup, Richard F. Nash, and Alan M. Wagner. "The tonic immobility reaction in chickens: Response characteristics and methodology". In: *Behavior Research Methods & Instrumentation* 3.5 (1971), pp. 237–239. DOI: 10.3758/bf03208389.

[12] J.G. White et al. In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314.1165 (1986), pp. 1–340. DOI: 10.1098/rstb.1986.0056.

[13] Steven J. Cook et al. "Whole-animal connectomes of both caenorhabditis elegans sexes". In: *Nature* 571.7763 (July 2019), pp. 63–71. DOI: 10.1038/s41586-019-1352-7.

[14] Yung-Chi Huang et al. "A single neuron in C. elegans orchestrates multiple motor outputs through parallel modes of transmission". In: *bioRxiv* (2023). DOI: 10.1101/2023.04.02.532814. eprint: https://www.biorxiv.org/content/early/2023/04/02/2023.04.02.532814.full.pdf. URL: https://www.biorxiv.org/content/early/2023/04/02/2023.04.02.532814.

[15] Albert Gidon et al. "Dendritic action potentials and computation in human layer 2/3 cortical neurons". In: *Science* 367.6473 (2020), pp. 83–87. DOI: 10.1126/science.aax6239.

[16] Marc Hammarlund et al. "The CENGEN project: The Complete Gene Expression Map of an entire nervous system". In: *Neuron* 99.3 (2018), pp. 430–433. DOI: 10.1016/j.neuron.2018.07.042.

[17] Shira Sardi et al. "New types of experiments reveal that a neuron functions as multiple independent threshold units". In: *Scientific Reports* 7.1 (2017). DOI: 10.1038/s41598-017-18363-1.

[18] Jerry E Mellem et al. "Action potentials contribute to neuronal signaling in C. elegans". In: *Nature Neuroscience* 11.8 (2008), pp. 865–867. DOI: 10.1038/nn.2131.