EFFICIENT NEURAL STIMULATION WITH IN-SILICO SIMULATION

Martin Schrimpf
McGovern Institute for Brain Research
Department of Brain and Cognitive Sciences
Center for Brains, Minds and Machines
Massachusetts Institute of Technology

ABSTRACT

Despite recent progress in neurotechnology hardware for neural stimulation, precise behavioral control remains challenging. Part of the challenge stems from complex interactions at the network level that make it difficult to link cellular stimulation to behavior. At the same time, particular artificial neural networks have recently been shown to accurately predict neural firing rates across the ventral stream as well as image-by-image behavior. Here, we extend model-to-brain mappings with a stimulation module that translates from micro-stimulation in primate brains to its neural and behavioral effects in silico. We demonstrate the potential efficacy of this approach by reproducing psychometric shifts in face-selective sites observed by Afraz et al. (2006). Finally, we argue that the stimulation module can be used to efficiently determine the most effective stimulation patterns while accounting for complex network interactions, and make predictions on which changes to the neurotechnological hardware will be most impactful for biasing behavior.

1 INTRODUCTION

In recent years, neurotechnological hardware for controlling spike rates in neural tissue has made great progress with the rise of optogenetics (Bernstein and Boyden, 2011), more precise applications of microstimulation (Salzman et al., 1992; Romo et al., 1998; Romo and Salinas, 1999; Afraz et al., 2006; 2015), or inactivation techniques like pharmacological muscimol injections (Arikan et al., 2002; Rajalingham and DiCarlo, 2019). In basic science, stimulation studies typically first carefully map out neuronal tissue to locally preferred stimulus categories (Schalk et al., 2017; Afraz et al., 2006) and then stimulate those locations in order to bias behavior in the direction of the stimulated

location's preferred category. Schalk et al. (2017) for example were able to evoke "facephenes" by stimulating the fusiform face area (FFA), following earlier efforts to find the anatomical location of functional face processing (Kanwisher et al., 1997).

Precisely controlling behavior through neural stimulation on the other hand has remained elusive. For instance, while we can evoke the general concept of a face, it is unclear how we would stimulate for *specific faces*. In part, this is due to the difficulty of tightly linking the effects of neural stimulation to behavioral responses. For instance, in optogenetics, even though tissue is meant to be activated through excitations of cells, network-wide suppression effects can occur that, instead of exciting the network, actually lead to reduced population activity, resulting in an inability to impact behavior. These effects can be attributed to network effects (Jazayeri and Afraz, 2017; Dayan et al., 2013) where complex interactions in the network lead to counter-intuitive behavior of the neural population and as a result, counter-intuitive behavior of the organism as a whole.

As such, clinical applications are limited to applications of coarse neural stimulation. For instance, Deep Brain Stimulation (DBS) statically applies coarse electric pulses in order to drive neural activity in the region the electrode is implanted in (Ashkan et al., 2017). Due to the coarse application of stimulation, clinical applications have so far focused on diseases where stimulating relatively large parts of a brain area leads to improved behavior, such as Parkinson, mood disorders, and epilepsy. In these settings, increasing the activity of an area is sufficient to improve the patients' behavior. However, coarse stimulation techniques fall short when increasing an area's activity is not sufficient to cure behavioral deficits: for instance, deficits in the visual cortex can stem from lesions in visual cortex (van Polanen and Davare, 2015) where part of an area, and thus its activity, are missing. In this case, we would have to *precisely reenact* the activity of the damaged neurons which requires us to account for the complex network effects described earlier.

To capture complex network effects, we here turn to artificial neural networks (ANNs) which currently constitute the most accurate predictive models of complex activity in the primate visual stream and the object recognition behavior it supports. Specifically, evoked internal representations of specific ANNs are remarkably similar to evoked representations in V1, V2, V4, and inferior temporal (IT) cortex (Yamins et al., 2014; Cadena et al., 2017; Schrimpf et al., 2018; 2019; Kar et al., 2019). ANNs have further been shown to accurately predict primate behavior, to some extent even on an image-by-image level (Rajalingham et al., 2018; Schrimpf et al., 2018). These evaluations are usually carried out by showing the ANN the same images as are shown to primates, and then comparing whether e.g. the ANN's predicted IT_{ANN} representations lie in the same linear subspace as recorded IT representations $\text{IT}_{\text{recording}}$, i.e. whether a linear regression from IT_{ANN} to $\text{IT}_{\text{recording}}$ can predict held-out recordings. For behavior, the ANN's response error patterns are compared to the error pat-

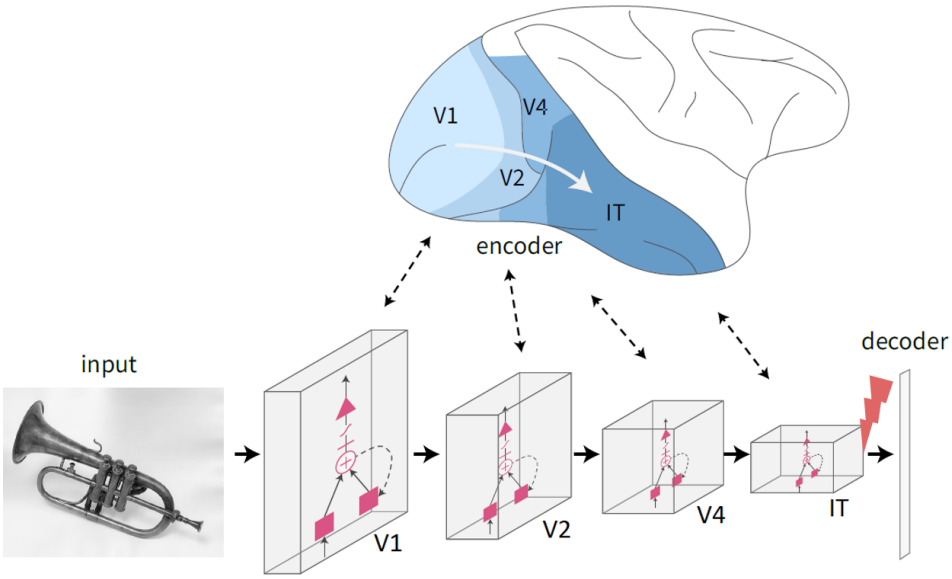

Figure 1: **Brain to model comparisons and the proposed stimulation intervention.** Artificial Neural Networks make predictions on evoked firing rates and behavior which have been shown to correspond well to neural recordings and behavioral measurements in primates. We here aim to extend this approach with a stimulation module (red lightning) that can be added onto existing models.

terns by human subjects (see Figure 1 for an overview). Recently, ANNs have also been applied in a neurotechnology context, where neural populations are non-invasively controlled by synthesized images that are generated through the model (Bashivan et al., 2019).

Here, we leverage artificial neural networks in order to simulate *in silico* the behavioral effects of neural stimulation *in vivo*. In particular, we propose a stimulation module that can be added onto any model that makes predictions on neurons and behavior, and demonstrate its usefulness by reproducing psychometric shift effects following stimulation observed by Afraz et al. (2006). We then go beyond the data and simulate many different stimulation patterns for which the model predicts strong behavioral effects. If correct, these precise stimulation patterns would yield much more efficient biasing of behavior that goes beyond the effects from the coarsely chosen stimulation patterns used so far. Since we have full control over the stimulation module, we can also evaluate the impact of potential future neurotechnological hardware. For the specific task of biasing face responses in IT, we determine that the activation of local, isolated sites together with a stronger excitation of neural firing rates, would lead to the greatest ability of biasing behavioral responses. These predictions might help guide neuroscience tool development in a direction that is most impactful for stimulated behavioral effects. On the other hand, model predictions might very well turn out to be incorrect in which case this specific model (specific ANN + specific stimulation module) have been falsified, and need to be updated – a step that leads to progress in obtaining better and better models.

2 TRANSLATING MODEL ACTIVATIONS TO ON-TISSUE FIRING RATES

We start with "CORnet-Z", an anatomically mapped model of brain processing in the ventral stream, with weights pre-trained on ImageNet (Kubilius et al., 2018). In order to tie the predictions of neural firing rates in IT to physical tissue, we use Utah array recordings from Majaj et al. (2015): the model is shown the same 2,560 stimuli that evoked the monkeys' neural firing rates y, we "record" the model's IT representations x, and construct a linear transformation that maps from model to monkey representations,

$$y = \mathbf{W}x \tag{1}$$

where $\mathbf{W}$ denotes linear regression weights learned with PLS regression. Notably, since the monkey IT recordings stem from an electrode array implanted in physical tissue, predicted firing rates $\hat{y}$ are grounded in a tissue map where each site has neighbors in a 2D space along the Utah electrode array. After this initial mapping step, we can now simulate IT firing rates using the model and the linear transformation, without relying on costly monkey recordings.

3 GAUSSIAN-SCALED ADDITIVE STIMULATION

After obtaining a tissue-mapped predictive model of firing rates, we here outline the stimulation module that models the effects of stimulation on the model. Previous literature has started to uncover the effects of microstimulation on the neural sites themselves: largely, responses seem to be driven most at the center of stimulation, with an increasing falloff as the distance from the center increases (Tolias et al., 2005; Lee et al., 2014). Here, we model this effect with an additive 2D Gaussian that is positioned at the center of stimulation, parametrized by its covariance σ^2. The second free parameter is the scaling multiplier λ that translates from e.g. 50 mA microstimulation to an additive firing rate at the center of the Gaussian. The full effect of the stimulation module on firing rates $\hat{y}$ at a position pos is thus

$$\hat{y}'_{pos} = \hat{y} + \lambda \times \mathcal{N}(pos|\mu, \sigma^2) \tag{2}$$

where μ is the 2D center of stimulation in coordinates of the Utah array (see Figure 3 (left) for an example).

Note that a potential confound in this modeling choice is that nearby neurons in the original recordings could have only gotten activated due to lateral connections from the stimulated neuron itself. However, since the model used in this study does not capture these lateral connections, the spatial spread of stimulation might compensate for that shortcoming.

4 REPRODUCING PSYCHOMETRIC SHIFTS FOLLOWING STIMULATION IN FACE-SELECTIVE SITES

In the following, we attempt to reproduce behavioral effects following microstimulation, observed by Afraz et al. (2006). Specifically, we aim to reproduce, in the model, 1) a behavioral shift towards more likely responding "face" over "no face" when stimulated; and 2) the stimulation of face-selective sites leading to a stronger psychometric shift compared to sites that are not face-selective (see Figure 2).

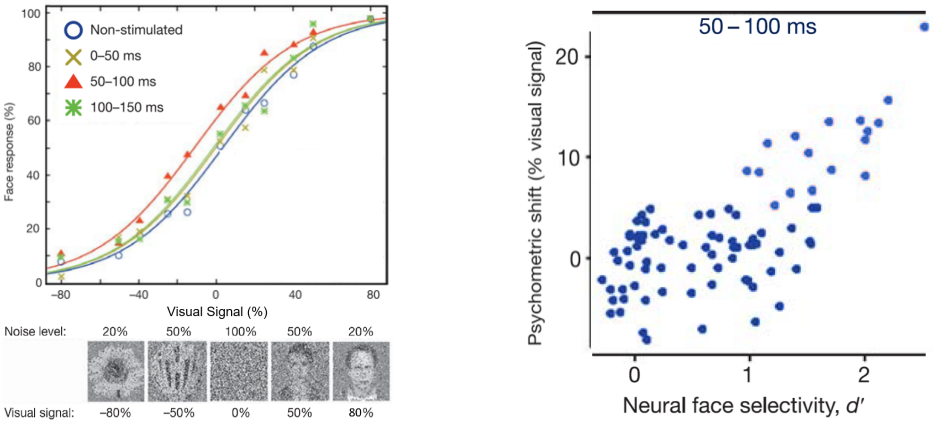

Figure 2: **Main findings by Afraz et al. (2006).** (*left, bottom*) Macaque monkeys are presented with face and non-face images with varying amounts of noise in a face categorization task. (*left, top*) Sample effect of stimulation on behavioral responses. Compared to no stimulation (blue line), stimulation during 50-100 ms (red line) leads to a shift in behavioral responses towards faces. (*right*) When stimulating sites that are more face-selective, the elicited psychometric shift is pushed more towards faces compared to sites that are not face-selective.

Since we don't have access to the original stimuli, we synthesized our own: for the face images, we randomly select 500 images from the labeled-faces-in-the-wild dataset (Huang et al., 2012). For the non-face images, we chose 6 Imagenet (Deng et al., 2009) categories that roughly follow the data description by Afraz et al. (2006) (folding chair, pineapple, hen, sea cucumber, car wheel, mountain bike) from which we randomly select 500 images. For each of the total 1,000 images, we then randomly choose a noise level between $0 - 100\%$ and distort the image with random gray-scale values between $0 - 255$.

To connect neurons to the task of face/no-face discrimination in Afraz et al. (2006), we train a binary logistic decoder that, based on the IT representations $\hat{y}$, distinguishes images into "face" and "no face". We train the decoder on 500 randomly selected images and perform all further testing on the held-out 500 images. Using this training procedure, we are able to reproduce the baseline behavioral response (Figure 2, left, blue curve) in the model (Figure 3, middle, blue curve).

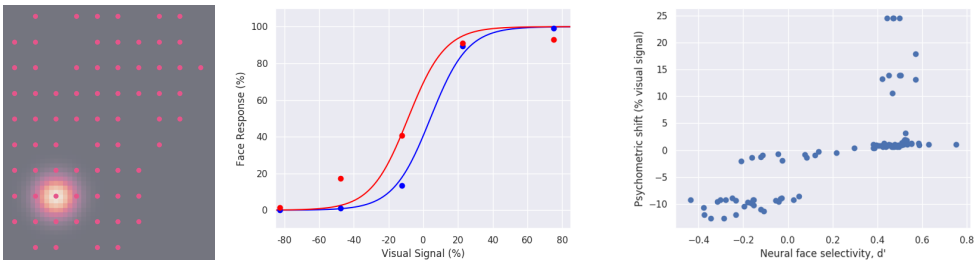

Figure 3: **Model stimulation qualitatively reproduces behavioral effects.** (*left*) Spatial application of a model stimulation (orange blob, covariance $\sigma^2 = 0.1$) to the location on a Utah array (red dots). (*middle*) Sample effect of stimulation on behavioral responses. Compared to no stimulation (blue line), stimulation (red line) leads to a shift in behavioral responses towards faces. (*right*) Face-selective sites lead to a stronger bias in behavioral responses towards faces than sites that are not face-selective.

We instantiate our in-silico stimulation module with a spatial coverage of $\sigma^2 = 0.1$ and a mA translation of $\lambda = 0.01$. In each experiment, we randomly choose a neural site, center the Gaussian stimulation on that point and add a spatially scaled stimulation vector to the IT firing rates $\hat{y}$ as described in Section 3. In the stimulation case, the decoder then receives the stimulated IT responses $\hat{y}'$ which lead to changes in the model's behavior. Figure 3 (middle, red curve) shows how for a sample site in the model, we can reproduce the behavioral shift of the psychometric function towards faces (compare with Figure 2, left, red curve). Aggregating across all the sites, we find a similar effect to what Afraz et al. (2006) reported: sites that are more face-selective lead to a greater behavioral bias towards faces than sites that are not face-selective. Note that overall, less of the sites here are face-selective which most likely stems from the dataset we used (Majaj et al., 2015) not being collected with face-selectivity in mind.

5 Predictions outside the data

On top of capturing existing effects, we can also further simulate the in-silico model and quickly explore the impact of possible experiments. One central promise of computational models of the brain that capture complex interactions is to precisely predict behavioral effects. In the case of stimulation, this would entail using the model to find precise stimulation patterns that maximally push behavior in the desired direction, where simpler models (e.g. word models) fail to precisely capture the complex interactions from neurons to behavior. For instance, in optogenetics, despite activating part of the neurons, overall network suppression effects can occur that are not captured by simple mechanic explanations (Jazayeri and Afraz, 2017), but could be predicted by computational models. Figure 4 outlines this upshot of using computational models: first, strong models should make more precise predictions on experimental outcomes and second, they are likely to be more efficient (Yamins and DiCarlo, 2016) in finding stimulation patterns that maximally drive behavior.

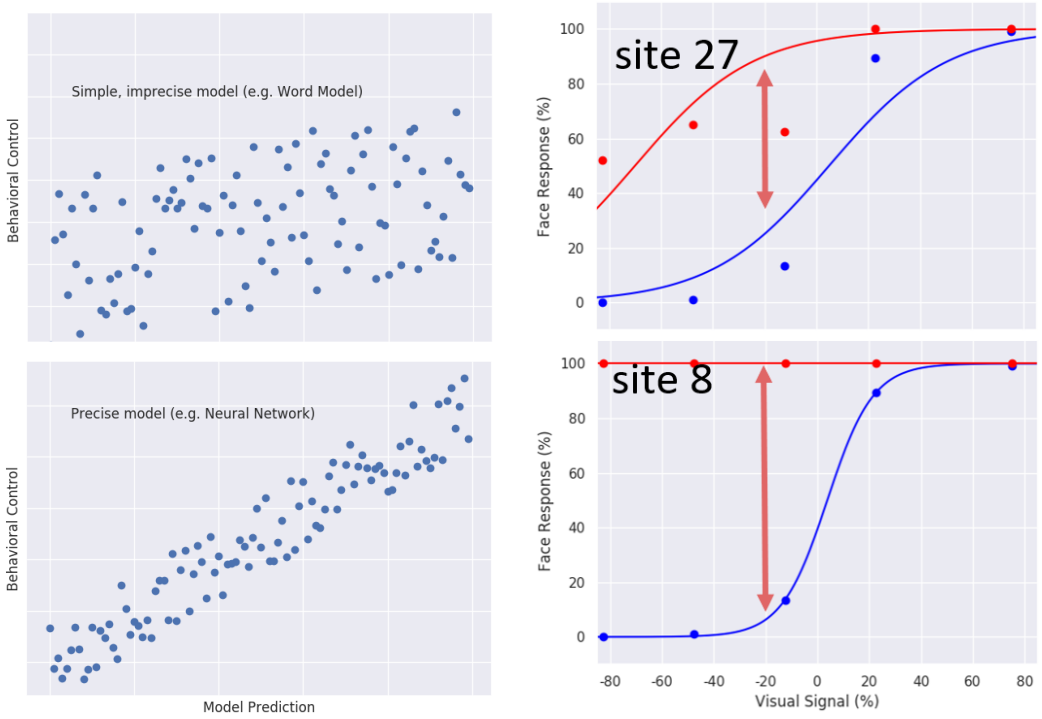

Figure 4: **Potential stronger behavioral control through model-driven stimulation.** (*left*) Sketched hypotheses of how an efficient, and more precise model of network interactions increases behavioral control compared to an imprecise model lacking complex network interactions. (*right*) Sample sites that the model predicts to lead to stronger behavioral bias towards faces.

The two major reasons for this efficiency gain are the ability to execute the in-silico model much more quickly than an animal model, and the transparency of the in-silico system which allows to backtrack changes through the entire model (Bashivan et al., 2019). Given the model's predictive power and efficiency, we can easily query it for stimulation patterns that would e.g. bias primates to faces altogether.

On top of utilizing the model for determining the most impactful stimulation patterns, we can also play out the effect of different changes to neurotechnology hardware. Recall for instance that the stimulation's spatial coverage σ^2 and the scaling multiplier λ mapping onto firing rates have been chosen to match the Afraz et al. (2006) paper. We can change these parameters in order to gauge how e.g. more spatially isolated microstimulation hardware would impact behavior.

Figure 5 illustrates this analysis: each subplot shows the psychometric shift when stimulating sites of varying face-selectivity (same notation as in Figure 3, right), the columns denote increasing spatial coverage, and the rows increasing excitation of firing rates. The circled subplot reflects the parameter choices that best mapped onto Afraz et al. (2006). In this analysis of different parameter combinations, the neurotechnology hardware to most strongly allow pushing behavior maximally apart between face and non-face responses, would stimulate only local isolated sites and drive those

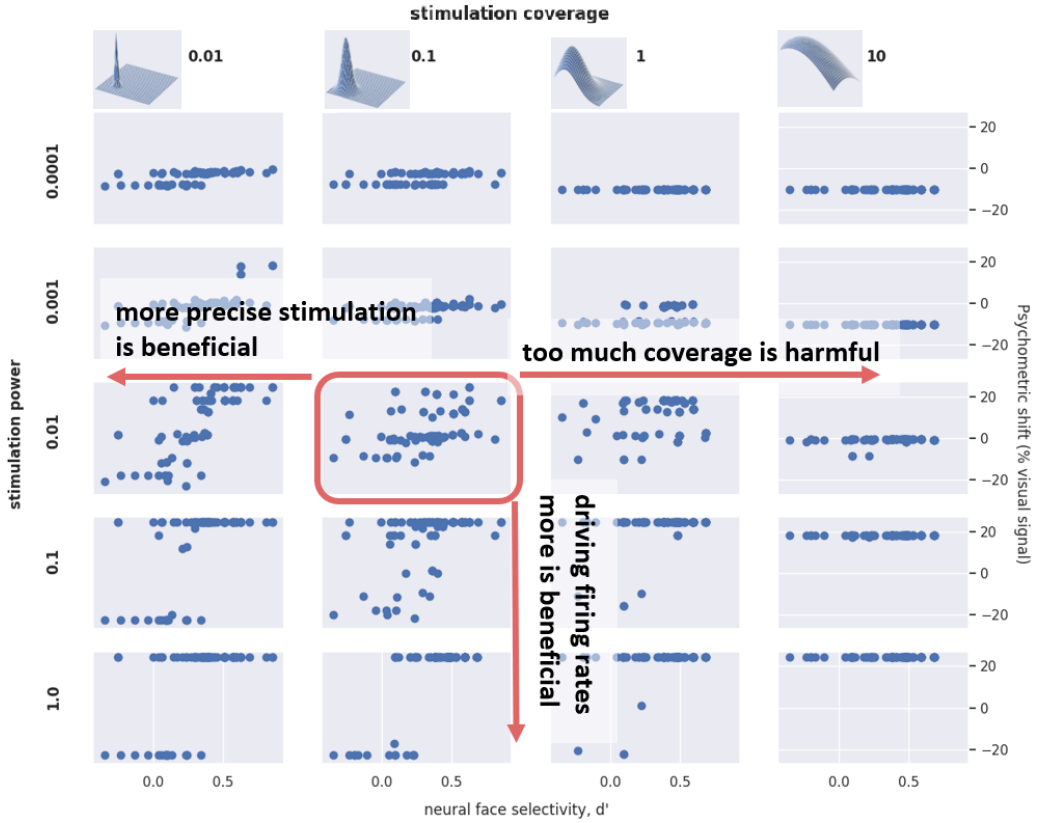

Figure 5: **Predictions on neurotechnology hardware.** Model simulation suggests that neurotechnological hardware that stimulates units in an isolated manner, and with stronger excitation of firing rates, will be able to maximally tease face/non-face responses apart.

sites' firing rates very high. These are however very preliminary results with many assumptions: the model and stimulation module have only been tested qualitatively on a single experimental paradigm, despite isolated spatial stimulation local interactions could still distort the signal, and firing rates cannot be driven infinitely high.

6 Next Steps

All of these prototypical analysis are meant to show-case the potential benefits of an in-silico stimulation model onto neurotechnology hardware. I am personally hoping to use these initial experiments as a stepping stone for in-vivo experiments in macaques. By having control over both recordings and stimulation, I could optimize mapping parameters (such as stimulation coverage σ^2 and excitation multiplier λ) for the animal model they will later be used on, and test model predictions in the loop. A first step might be to capture the effects of e.g. V4 microstimulation on IT recordings, before making the larger leap to precise behavioral control. For behavior, the grand goal would be to predictably control visual perception and strongly bias behavioral responses. For instance, a monkey

would look at a blank screen, we would query the model and ask how we should stimulate IT in order to yield a certain behavioral response of e.g. a dog in a match-to-sample task, then conduct that stimulation on the monkey, record the monkey's response, and evaluate whether the stimulation led to the desired behavior.

In the long run, another hope is to apply efficient model-driven stimulation in a clinical setting to tackle certain psychiatric diseases. For instance, patients with ventral stream lesions could have a stimulation implant reenacting the activity that is missing due to the lesion. If inducing behavioral percepts should work out, this approach could be expanded to other perceptual domains, and perhaps even to cognitive ones such as language understanding. On top of that, after detecting schizophrenic delusions, we could potentially even use this tool to create stimulation patterns that undo the delusional episode. Since the model can predict perception from neural activity, we can find a neural activity pattern that would reset perception back to normal. Through the stimulation module proposed here, we could then apply the computationally determined stimulation pattern to the patient which – given a correct model – would undo the delusion. Since the stimulation module is only an add-on onto a computational brain model, there could further be different instantiations of stimulation such as an optogenetic module, or a magnetic-stimulation module (Chen et al., 2015). In principle, the approach would also work across species, given a computational model of the species' brain processing.

All of these experiments would exemplify causal control over the brain through a computational simulation of the brain. Without the in-silico model, we would have to semi-randomly search for the right stimulation pattern directly in-vivo which is much less efficient. The simulated model stimulation allows us to quickly determine the stimulation pattern that is most likely to elicit the desired behavioral effect.

REFERENCES

Seyed Reza Afraz, Roozbeh Kiani, and Hossein Esteky. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 2006.

Jacob G. Bernstein and Edward S. Boyden. Optogenetic tools for analyzing the neural circuits of behavior. *Trends in Cognitive Sciences*, 15(12):592–600, 2011.

C D Salzman, C M Murasugi, K H Britten, and W T Newsome. Microstimulation in visual area MT: effects on direction discrimination performance. *Journal of Neuroscience*, 12(6):2331–2355, 1992.

Ranulfo Romo, Adrián Hernández, Antonio Zainos, and Emilio Salinas. Somatosensory discrimination based on cortical microstimulation. *Nature*, 392(6674):387–390, 1998.

Ranulfo Romo and Emilio Salinas. Sensing and deciding in the somatosensory system. *Current Opinion in Neurobiology*, 9(4):487–493, 1999.

Arash Afraz, Edward S. Boyden, and James J. DiCarlo. Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences (PNAS)*, 112(21):6730–6735, 2015.

Rasim Arikan, Nicquet M.J Blake, Joseph P Erinjeri, Thomas A Woolsey, Lisette Giraud, and Stephen M Highstein. A method to measure the effective spread of focally injected muscimol into the central nervous system with electrophysiology and light microscopy. *Journal of Neuroscience Methods*, 118(1):51–57, 2002.

Rishi Rajalingham and James J. DiCarlo. Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits. *Neuron*, 102(2):493–505, 2019.

Gerwin Schalk, Christoph Kapeller, Christoph Guger, Hiroshi Ogawa, Satoru Hiroshima, Rosa Lafer-Sousa, Zeynep M. Saygin, Kyousuke Kamada, and Nancy Kanwisher. Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proceedings of the National Academy of Sciences (PNAS)*, 2017.

N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–11, 1997.

Mehrdad Jazayeri and Arash Afraz. Navigating the Neural Space in Search of the Neural Code. *Neuron*, 93(5): 1003–1014, 2017.

Eran Dayan, Nitzan Censor, Ethan R Buch, Marco Sandrini, and Leonardo G Cohen. Noninvasive brain stimulation: From physiology to network dynamics and back. *Nature Neuroscience*, 16(7):838–844, 2013.

Keyoumars Ashkan, Priya Rogers, Hagai Bergman, and Ismail Ughratdar. Insights into the mechanisms of deep brain stimulation. *Nature Reviews Neurology*, 13(9):548–554, 2017.

Vonne van Polanen and Marco Davare. Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, 79:186–191, 2015.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv preprint*, 2017.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J Dicarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv preprint*, 2018.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Corey Ziemba, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James DiCarlo. Using Brain-Score to Evaluate and Build Neural Networks for Brain-Like Object Recognition. In *Cosyne*, 2019.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 2019.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 2018.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J Dicarlo. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, 2018.

Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

Andreas S. Tolias, Fahad Sultan, Mark Augath, Axel Oeltermann, Edward J. Tehovnik, Peter H. Schiller, and Nikos K. Logothetis. Mapping cortical activity elicited with electrical microstimulation using fMRI in the macaque. *Neuron*, 48(6):901–911, 2005.

Taekwan Lee, Lili X Cai, Victor S Lelyveld, Aviad Hai, and Alan Jasanoff. Molecular-level functional magnetic resonance imaging of dopaminergic signaling. *Science*, 344(6183):533–535, 2014.

Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Ritchie Chen, Gabriela Romero, Michael G Christiansen, Alan Mohr, and Polina Anikeeva. Wireless magnetothermal deep brain stimulation. *Science*, 347(6229):1477–1480, 2015.