Characterizing and predicting cortical evoked responses to direct electrical stimulation of the human brain

Abbreviated title: Predicting responses to cortical stimulation

Cynthia Steinhardt[1], Pierre Sacre[1], Timothy C. Sheehan[2], John H. Wittig, Jr [2], Sara K. Inati [3], Sridevi Sarma[1], Kareem A. Zaghloul[2 †]

[1] Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218

[2] Surgical Neurology Branch, NINDS, National Institutes of Health, Bethesda, MD 20892

[3] Office of the Clinical Director, NINDS, National Institutes of Health, Bethesda, MD 20892

[†]**Correspondence should be addressed to:**
Kareem A. Zaghloul
Surgical Neurology Branch, NINDS, National Institutes of Health Building 10, Room 3D20
10 Center Drive Bethesda, MD 20892-1414
Office: (301) 496-2921
Email: kareem.zaghloul@nih.gov

Draft Date: March 27, 2020
Number of Figures: 4
Number of Supplementary Figures: 5

Highlights

- Cortical evoked responses can be characterized by a linear system.
- Generated linear models predict the responses to individual stimulation pulses.
- Models also predict the responses to combinations of stimulation pulses.
- This offers an approach for predicting responses to more complex stimulation.

Abstract

Background: Direct electrical stimulation of the human brain has been used to successfully treat several neurological disorders, but the precise effects of stimulation on neural activity are poorly understood. Characterizing the neural response to stimulation, however, could allow clinicians and researchers to more accurately predict neural responses, which could in turn lead to more effective stimulation for treatment and to fundamental knowledge regarding neural function.

Objective: Here we use a linear systems approach in order to characterize the response to electrical stimulation across cortical locations and then to predict the responses to novel inputs.

Methods: We use intracranial electrodes to directly stimulate the human brain with single pulses of stimulation using amplitudes drawn from a random distribution. Based on the evoked responses, we generate a simple model capturing the characteristic response to stimulation at each cortical site.

Results: We find that the variable dynamics of the evoked response across cortical locations can be captured using the same simple architecture, a linear time-invariant system that operates separately on positive and negative input pulses of stimulation. We demonstrate that characterizing the response to stimulation using this simple and tractable model of evoked responses enables us to predict the responses to subsequent stimulation with single pulses with novel amplitudes, and the compound response to stimulation with multiple pulses.

Conclusion: Our data suggest that characterizing the response to stimulation in an approximately linear manner can provide a powerful and principled approach for predicting the response to direct electrical stimulation.

Keywords electrical brain stimulation, human cortex, cortical evoked response, linear system identification

Introduction

Clinicians and researchers have been directly stimulating the human brain with electrical current for decades in order to address a variety of neurological disorders and to gain insight into the neural circuits that govern function and behavior [1,2]. Yet despite this vast experience, precisely how electrical stimulation affects neural circuitry and neuronal activity in the human brain remains unclear [2, 3]. Direct electrical stimulation depolarizes membrane potentials, leading to neural responses both locally and in distant neural structures through activation of axons underlying the stimulation site [4–6]. The distal effects of stimulation are complex and involve direct activation of orthodromic, and to a lesser extent antidromic, propagation pathways to the target site as well as indirect activation of local circuits within those target regions [7,8].

These complex local and distal effects have therefore made it difficult to anticipate and predict the neural responses to stimulation. This has consequently presented a challenge for accurately interpreting the behavioral effects of electrical stimulation when used routinely for clinical purposes or for research. Stimulation can result in positive responses, such as evoked phosphenes or motor movements [9, 10], but stimulation can also induce negative responses such as in speech arrest [11]. Similarly, stimulation can have variable effects on higher order cognitive functions such as memory [12–14]. Stimulation at a single site can even promote both positive and negative responses, likely depending on the broader regions that are connected to and therefore targeted by activation of the stimulation site [2, 11, 15]. These effects of stimulation suggest that there is a need to well characterize the responses to stimulation across broader brain regions.

There are clear benefits to well characterizing the effects of stimulation on neural activity. First, bettter understanding of the effects of stimulation could allow clinicians and researchers to more accurately predict neural responses. This in turn could lead to developing more effective stimulation algorithms for treating neurological disorders that are more adaptive to the individual patient, that require less power, and that produce fewer side effects. Second, characterizing the effects of stimulation may provide fundamental knowledge regarding neural function. By directly manipulating neural responses in a controlled manner, stimulation may be used to provide insight into neural structure and connectivity between different brain regions and to investigate the causal role of neural activity on behavior. Direct stimulation with individual current pulses that result in cortical evoked potentials has been successfully used, for example, to map connections between brain regions [16–18].

Here, we demonstrate that by directly delivering single pulses of electrical stimulation to the human brain with amplitudes drawn from a random distribution, we can characterize and subsequently predict the responses in other brain regions to stimulation at a given site. Although the dynamics of the evoked response to stimulation are governed by the variable interconnections between each brain region and the stimulation site, we find that the responses in each brain region can be characterized using the same simple architecture - a linear time-invariant system that operates separately on positive and negative input pulses of stimulation. Characterizing the responses to stimulation using this simple and tractable model enables us to accurately predict the response to subsequent stimulation using individual pulses with novel amplitudes, and the compound response to stimulation using combinations of multiple pulses. Our results therefore suggest that the complex mechanisms that underlie the effects of stimulation of one site on another can be captured in an approximately linear manner. Moreover, our results suggest an approach for building more complex electrical stimuli and anticipating the responses to stimulation in other brain regions.

Material and Methods

Ten individuals (8 male; 32.5 ± 0.9 years) with drug resistant epilepsy underwent a surgical procedure in which subdural platinum recording contacts (3 mm exposed diameter; PMT Corporation, Chanhassan MN) were implanted on the cortical surface. In all cases, placement of the contacts was determined by the clinical team in order to best localize epileptogenic regions for resection. Data were collected at the Clinical Center at the National Institutes of Health (NIH; Bethesda, MD). The research protocol was approved by the Institutional Review Board, and informed consent was obtained from all participants in the study. All data are reported as mean ± SEM unless otherwise noted.

In order to characterize the responses to electrical stimulation, we applied individual biphasic pulses to a pair of neighboring electrodes while recording the intracranial EEG (iEEG; Nihon Kohden Inc., Irvine CA) responses in the remaining electrodes in each participant. For each individual pulse, we used a stimulation amplitude that was randomly drawn from a uniform distribution of amplitudes between approximately 8 and -8 mA (Fig. 1C). We presented trials of single pulses of electrical stimulation in two experimental sessions while capturing iEEG recordings in each participant (Fig. 1F). In each session, trials were separated by approximately one second, resulting in a stimulation frequency of approximately 1 Hz. We used the first session, which we refer to as training, in order to generate a tractable model that characterizes how each recording site responds to stimulation at each stimulation site. During the second session, which we refer to as testing, we used trials of individual pulses, consisting of novel stimulation amplitudes also drawn randomly from the same distribution, and trials containing multiple pulses to test how well the identified model can predict the evoked response to novel stimulation.

We identified every electrode that was responsive to stimulation (Supplementary Fig. S1). For each responsive electrode, we constructed a model that characterizes the evoked response to stimulation at the stimulation site by examining the evoked responses to the input sequence of pulses presented during the five-minute training session (see Suppplementary Material and Methods). Each model had an identical architecture, comprised of a linear system with impulse responses, $h^+[t]$ and $h^-[t]$, that separately operate on positive and negative input pulses, respectively. Hence, given an input, $x[t]$, our model estimates a predicted output, $\hat{y}[t] = \mathcal{H}\{x[t]\}$ by separately convolving the positive and negative input pulses with the impulse responses, $h^+[t]$ and $h^-[t]$, respectively:

$$\hat{y}[t] = \mathcal{H}\{x[t]\} = \sum_{\tau=0}^{L} h^+[\tau]f^+(x[t-\tau]) + \sum_{\tau=0}^{L} h^-[\tau]f^-(x[t-\tau]),$$

where $f^+(x)$ and $f^-(x)$ are rectifiers that select positive and negative pulses, respectively, and where each impulse response is of finite duration, L. The model $\mathcal{H}$ is therefore fully parametrized by the two finite impulse responses, $h^+[t]$ and $h^-[t]$.

Using the derived impulses responses, $h^+[t]$ and $h^-[t]$, we then estimated a predicted output, $\hat{y[t]}$, to the pulses presented during the testing experimental session. We computed the prediction error between $\hat{y}[t]$ and the observed output $y[t]$ by calculating the mean squared error (MSE) directly on the differences between their time series (see Supplementary Material and Methods).

Data availability

Processed data and MATLAB code used for analysis is available upon request.

Results

We collected intracranial EEG (iEEG) data from ten participants with drug resistant epilepsy who were being monitored for seizures with surgically implanted subdural electrodes. We recorded iEEG signals at every electrode while delivering biphasic current pulses to one site in the brain (Fig. 1A; Material and Methods). At each stimulation site, we used bipolar stimulation across two adjacent electrode contacts (Fig. 1B). We delivered one pulse approximately once per second, with the amplitude of each pulse drawn randomly from a uniform distribution of amplitudes between 8 mA and -8 mA (Fig. 1C). We considered the presentation of each individual stimulation pulse, followed by its inter-stimulus interval, a single stimulation trial. The stimulation amplitude of each pulse in each trial defines both phases of the biphasic pulse such that stimulation using a positive (negative) current amplitude resulted in an initial positive (negative) deflection followed by an equally sized negative (positive) deflection. The series of stimulation amplitudes has the characteristics of a white noise input since the amplitudes are uncorrelated and have zero mean.

At each unique stimulation site ($n = 12$ across participants), we delivered a five minute training session consisting of trials in which one pulse was delivered at the beginning of each trial, and with each trial and therefore each pulse delivered approximately once every second, resulting in a stimulation frequency of approximately 1 Hz. We examined the evoked responses captured on the iEEG traces at every other electrode to every trial of single pulse stimulation during training (Fig. 1D). In many example electrodes, stimulation resulted in an evoked response that appeared to increase in magnitude with stimulation amplitude (Fig. 1E). We often observed this increase in the evoked response to both positive and negative stimulation. Based on the average evoked response to all stimulation trials, we determined whether each electrode in each participant was responsive to stimulation (Material and Methods; Fig. 1B; Supplementary Figs. S1 and S2). We examined only the electrodes that exhibited an average evoked response to stimulation that met our criteria for responsiveness in subsequent analyses (52 total responsive electrodes, 4.3 ± 1.42 electrodes, corresponding to $5.02 \pm 1.7\%$, per participant; for analyses of remaining electrodes, see Supplementary Figures).

At each stimulation site, we subsequently presented a two-minute test series containing trials of individual pulses, with each trial and therefore each pulse also delivered with a stimulation frequency of approximately 1 Hz. The trials used in the test series were comprised of individual pulses with novel stimulation amplitudes also drawn from a uniform distribution of amplitudes between 8 mA and -8 mA (Fig. 1F). Our initial goal was to determine whether we could use the evoked responses captured during training to generate a simple linear model capable of predicting the responses to novel individual pulses of stimuli during testing. We reasoned that if we could generate such a model, we could then use the same linear model to predict the compound response to multiple pulses delivered in quick succession.

Consistent Responses Across Time

One requirement for being able to successfully predict responses to direct cortical stimulation is that the response to the same stimulus should be consistent across time, or time-invariant. We examined this by repeating the two-minute test series of trials of single pulses between three and five times in each participant (Material and Methods). We observed that the evoked responses to the same pulses were indeed consistent (Fig. 2A). To quantify this consistency, we computed the average response across all but the first repetition of the same two-minute test series of trials of pulses and examined how well that average response trace could be used to predict the responses observed during the first presentation of the test series of trials (Fig. 2A). We used the residual difference between that average response trace and the response observed during the first presentation of the same trials of pulses to compute the mean squared error (MSE; Material and Methods). This measure, the MSE between the predicted and observed trace, reflects the extent to which the brain's response to stimulation is consistent across time. Across responsive electrodes, we found that the average MSE was $0.024 \pm 0.003\text{mV}^2$ (Fig. 2B). We also computed the average response across all but the final repetition, and similarly found that the MSE between the predicted and observed trace in that final repetition was $0.027 \pm 0.003\text{mV}^2$.

We compared this MSE to the mean of the squares of each electrode's evoked response observed during the first presentation of the two-minute test series of trials, which is proportional to the energy of the evoked trace. This is equivalent to the MSE that would arise had we predicted a response that was the mean value of the recorded trace, which was zero, for all time points, and is the upper limit of any measure of MSE since this would represent the error we would observe had we made no prediction. We refer to this upper bound as $\text{MS}_{\text{signal}}$. Any meaningful predictions would necessarily have to result in a residual error that is lower than this upper bound. The MSE between the prediction using the average response and the observed response during that first repetition was significantly less than $\text{MS}_{\text{signal}}$ across all responsive electrodes ($0.024 \pm 0.003\text{mV}^2$ versus $0.057 \pm 0.014\text{mV}^2$, $t(51) = -2.65$, $p = .011$, paired t-test; Fig. 2B; Supplementary Fig. S3). Similarly, the MSE between the prediction using the average response and the observed response during the final repetition was significantly less than $\text{MS}_{\text{signal}}$ across all responsive electrodes ($0.027 \pm 0.003\text{mV}^2$ versus $0.060 \pm 0.014\text{mV}^2$, $t(51) = -2.70$, $p = .009$, paired t-test). Hence, the average response across multiple repetitions of the same series of trials of test pulses can be used to predict a significant portion of the response observed in a repeated presentation of that same series of trials of pulses.

We also compared the MSE between the predicted and observed response to the mean of the squares of each electrode's baseline activity at rest, which is proportional to that electrode's baseline energy. This is the lower limit of any measure of MSE, given a baseline level of noise on each electrode, and is the best level of error one can achieve with any prediction. We refer to this value as the MS_{noise}. The MSE of the prediction was significantly larger than this lower limit, MS_{noise} ($0.024 \pm 0.003\text{mV}^2$ versus $0.018 \pm 0.003\text{mV}^2$, $t(51) = 4.95$, $p < .001$). Together, these data suggest that the evoked responses to stimulation are relatively, but not strictly, consistent across time.

Prediction of Single Pulse Responses

We were interested in characterizing the response to stimulation across all recording sites in a simple and tractable model that can be used to predict future responses. For each responsive electrode, we therefore constructed a model with an identical architecture. Each model characterized the response to stimulation using a linear impulse response that separately operates upon positive and negative stimulation pulses based on the observation that both positive and negative stimulation amplitudes often evoked similar responses (Material and Methods). The input to each model is therefore a static nonlinear rectification that selects for positive or negative stimulation pulses, respectively (Fig. 3A). Although we observed in some electrodes that the magnitude of the evoked responses could be different depending on whether stimulation was applied with a positive or negative pulse (Supplementary Fig. S2C), we set the gain for this nonlinear rectification to be the same for both positive and negative inputs. Because the evoked responses to stimulation appeared to scale with stimulation amplitude for each polarity of stimulation and were relatively consistent across time, the rectification is then followed by a linear time-invariant (LTI) operator.

A feature of the LTI operator is that we can derive the characteristic response to stimulation, or the impulse response, simply by minimizing the mean square error between the prediction any model would make and the observed evoked responses captured during the five minute training session (Fig. 3A,B). We therefore used the evoked responses to the five minute training session to derive the impulse response, $h[t]$, which fully characterizes the relation between an idealized input pulse at every cortical stimulation site and the anticipated response at every recording electrode site. Because we characterize the response to stimulation by separately processing positive and negative stimulation inputs, we derived a separate impulse response for positive and negative stimulation, $h^{+}[t]$ and $h^{-}[t]$ respectively, at every recording site (Fig. 3A).

We tested if we could use the simple models we generated that characterize the response to stimulation during training in order to predict the evoked response to the novel trials of single pulses presented during testing (Fig. 1F). One of the key assumptions in constructing a model using an LTI system is that the response amplitudes scale linearly with stimulation input. Examining the responses predicted by an impulse response function that operates upon single pulse inputs with different stimulation amplitudes, and com-paring the predicted responses to the responses observed during testing allows us to directly investigate this assumption. Using the derived impulse responses for each responsive electrode, we generated a predicted response, $\hat{y}[t]$, to the pulses presented during testing and that were comprised of novel stimulation ampli-tudes (Fig. 3C). As with the predictions based on the average response, we computed the MSE between the predicted output, $\hat{y}[t]$, and the observed output, $y[t]$. We compared this measure of error to the upper and lower bounds of error, $\mathrm{MS}_{\mathrm{signal}}$ and $\mathrm{MS}_{\mathrm{noise}}$ respectively. Across all responsive electrodes, the MSE using the predicted response, $\hat{y}[t]$, was significantly less than $\mathrm{MS}_{\mathrm{signal}}$ of the recorded response, $y[t]$ ($0.03 \pm 0.006\mathrm{mV}^2$ versus $0.057 \pm 0.014\mathrm{mV}^2$, $t(51) = -3.28$, $p = .0019$, paired t-test; Fig. 3D), but significantly greater than the lower bound, $\mathrm{MS}_{\mathrm{noise}}$ ($0.03 \pm 0.006\mathrm{mV}^2$ versus $0.018 \pm 0.003\mathrm{mV}^2$, $t(51) = 3.28$, $p = .0019$, paired t-test; Supplementary Fig. S4A). We found that this level of MSE of the predicted response was consistent when we examined all trials of different stimulation amplitudes, although the trials with the most negative stimulation amplitude exhibited slightly larger MSE (Supplementary Fig. S4B). In addition, we found that only approximately ten training trials were required in order to generate a linear model that was capable of predicting the evoked responses with a similar level of MSE as the model generated using all of the training trials (Supplementary Fig. S4C).

For every electrode, we quantified the percentage by which any prediction, $\hat{y}[t]$ reduced the error from the upper to lower bounds (Material and Methods). Given the baseline noise, the best prediction would reduce the error from the upper limit, $\mathrm{MS}_{\mathrm{signal}}$, proportional to the energy of the evoked response, to the minimum level, $\mathrm{MS}_{\mathrm{noise}}$, proportional to the energy of the noise, by 100%. Across responsive electrodes, we found that our predictions reduced the MSE by an average of $69.7 \pm 12.1\%$. Our model that separately rectifies positive and negative inputs before processing them with a linear time-invariant operator can therefore predict a significant component of the evoked response to a novel input.

We then compared how well the prediction based on the derived impulse responses for each brain region compared to the prediction based on the average responses. The MSE computed using the predicted response, $\hat{y}[t]$, approximately matched the MSE computed using the average of the repetitions for every responsive electrode (Fig. 3E; $t(51) = 1.38$, $p = 0.173$, paired t-test; Supplementary Fig. S4A). The similar levels of residual error suggest that our constructed model can predict responses to single pulses with novel input amplitudes with errors that are likely related to the consistency of the responses over time.

One concern with a predictive model that generates evoked responses that scale linearly with stimulation amplitude is that the observed evoked responses are not strictly linear (Fig. 1E). To confirm the predictive ability of our simple linear model, we therefore constructed a second model in which the initial rectification is replaced by a quadratic fit to the observed increases in signal energy that occur with increasing stimulation amplitude (Supplementary Fig. S4D). The input to this model is therefore a static nonlinear quadratic that operates on the input stimulation pulse amplitude, which is then followed by the linear time-invariant operator. We again calculated the MSE of the predicted output of this model when compared to the observed response during the testing session. Across all responsive electrodes, the MSE using the predicted response, $\hat{y}[t]$, was significantly less than the upper bound of error, $\mathrm{MS}_{\mathrm{signal}}$, of the recorded response, $y[t](0.029 \pm 0.006\mathrm{mV}^2$ versus $0.057 \pm 0.014\mathrm{mV}^2$, $t(51) = -3.36$, $p = .0015$, paired t-test), but significantly greater than the lower bound, $\mathrm{MS}_{\mathrm{noise}}$ ($0.029 \pm 0.006\mathrm{mV}^2$ versus $0.018 \pm 0.003\mathrm{mV}^2$, $t(51) = 3.06$, $p = .0036$, paired t-test; Supplementary Fig. S4D). The MSE of the predicted output using the quadratic input was not significantly different from the MSE using the linear model, suggesting that a linear model can account for much of the predicted response despite the apparent non-linear nature of the evoked response.

Prediction of Responses to Multiple Pulses

We were interested in whether characterizing the response to individual pulses of stimulation can also enable us to predict the compound response to combinations of multiple pulses. In this case, if multiple pulses are delivered before the response to any one pulse has time to return to rest, then the predicted response to these multiple pulses should simply be the superposition of the predicted responses to the individual pulses. This would suggest that this simple linear model operating separately on positive and negative stimulation pulses could be used to predict the responses to more complex patterns of electrical stimulation. To test this, in five participants who had a total of 22 responsive electrodes (4.4 ± 2.2 electrodes per participant), we delivered a series of trials during the experimental testing session in which some trials contained two to seven consecutive pulses. During these trials with multiple pulses, we presented each pulse within 20 to 50 ms of the previous pulse (Material and Methods). Hence, although each trial of stimulation was delivered approximately once every second, resulting in an overall stimulation frequency of approximately 1 Hz, within the trials with multiple pulses the stimulation frequency was not fixed. Within each trial with multiple pulses, any response evoked by the first pulse of this train of pulses would still contribute to the observed response when the second pulse was presented. In a typical example, the evoked responses to multiple consecutive stimulation pulses can be predicted by the model derived for this recording electrode (Fig. 4A).

We tested how well the predictions using the impulse responses derived for each electrode during training were able to predict the responses to multiple pulses presented during the two minute testing session. Across all responsive electrodes, the MSE using the predicted $\hat{y}[t]$ was significantly less than the upper bound of error, $\mathrm{MS}_{\mathrm{signal}}$, of the recorded output, $y[t]$ (0.048 ± 0.010 mV^2 versus 0.128 ± 0.041 mV^2, $t(21) = -2.31$, $p = .031$, paired t-test; Fig. 4B; Supplementary Fig. S5). We also compared the MSE of the prediction to the mean of squares of the baseline activity, $\mathrm{MS}_{\mathrm{noise}}$, which represents the lower limit of error that can be achieved with any prediction (0.048 ± 0.010 mV^2 versus 0.029 ± 0.005 mV^2, $t(21) = 2.86$, $p = .0093$). When predicting the response to multiple pulses, the predictions reduced the error by an average of 88.6 ± 22.9%.

As above, we also repeated the series of trials containing multiple pulses several times in order to compare the prediction of evoked responses using the model to the prediction we would obtain based on the average of several repetitions (Fig. 4C). In most cases, the model prediction matched the prediction using the average response. Across all responsive electrodes, the model prediction did not have a MSE that was significantly different than the MSE we found when using the average responses to predict the evoked response (0.048 ± 0.010 mV^2 versus 0.037 ± 0.006 mV^2, $t(21) = 1.65$, $p = 0.11$; Supplementary Fig. S5). Together, these data suggest that characterizing the response to stimulation using a linear impulse response that separately operates upon positive and negative inputs can predict the responses to multiple pulses, and suggest that the responses to more complex patterns of direct cortical stimulation can be estimated in an approximately linear manner.

Discussion

Our results demonstrate that by examining the responses to individual pulses with amplitudes drawn from a random distribution, we can characterize and subsequently predict the response to electrical stimulation across different brain regions. The evoked response in each electrode can be characterized using the same architecture, comprised of a linear system that separately operates on positive and negative input pulses. The information captured in these simple and tractable models has two important implications. First, characterizing neural responses to stimulation in this manner presents a powerful tool for predicting the future response to novel patterns of stimulation. Second, the features of each response may provide insight into underlying neural function and the structures and pathways connecting one region with another.

Direct stimulation of the brain has been utilized for several decades, but stimulation parameters generally have not been well explored even when applied to new disorders or patient populations [2]. In some of the most widely used and successful clinical applications, such as deep brain stimulation for movement disorders [23] or direct cortical stimulation mapping [11] for example, widely adopted stimulation parameters have largely been passed down based on previous studies or historical data. Stimulation often involves continuous or intermittent bursts of identical square wave pulses, and the main parameters that are manipulated relate to the location, frequency, amplitude, and width of the stimulating pulses.

Deviations from such standard stimulation algorithms, whether to improve treatment or to expand stimulation to new disorders or patient populations [24], rarely occur, largely because modeling and understanding all of the factors that underlie the various effects of stimulation has proven to be challenging [2, 3]. There are several reasons for this. Brain structures are interconnected in complex ways, and how information is propagated across this distributed network of neural circuits is still an active area of research [25, 26]. The fundamental units responsible for processing information within these networks, the neurons, are also complex, with electrophysiological dynamics that depend on neuron type, neuron location, interconnections with other neurons, and extra- and intracellular environments. Stimulation can affect these dynamics by changing the extracellular environment, which in turn can impact the activity of each neuron, ultimately leading to a network effect [27]. The underlying mechanisms for how this occurs are not fully understood. Finally, developing mechanistic and detailed models that can reliably take into account all of these complexities and predict neural responses to stimulation requires substantial data captured during stimulation from in vivo recordings of an entire neural circuit within an individual. These experiments are difficult to perform and must be done with care. Researchers have attempted to create realistic biophysical models that explicitly capture these dynamics, but these models are often high dimensional and nonlinear [28], and thus not tractable nor amenable, for example, to the design of closed-loop stimulation strategies that must operate in real time.

By characterizing the responses to stimulation using a simple and tractable model, however, and by demonstrating that these responses approximately add linearly, our data raise the possibility of addressing this challenge. Importantly, this approach solely relies upon the responses to stimulation itself and makes no assumptions about the underlying neurons or fibers of passage. Instead, we treat the brain as we would any other system whose inner mechanisms are relatively unknown but for which the inputs and outputs can be well controlled and described. We derive the relationship between stimulation and response based on the evoked responses to pulses of electrical stimulation. This approach builds upon previous work using single pulse electrical stimulation to investigate the evoked cortical responses [16–18, 29]. In this case, however, we use a random series of stimulation inputs with amplitudes that are temporally uncorrelated, an approach that has been successful in the study of mammalian vision in which random input pulses of light have been used to characterize how the retina and other higher order circuits process visual information [30, 31], as well as to characterize the effects of electrical stimulation on neural activity in the retina [32]. We use direct cortical stimulation with implanted electrodes to derive these responses here, but this approach is generalizable to any neuromodulation technology, including non-invasive stimulation modalities such as TMS or tDCS.

A relatively less explored, but equally important, use of direct stimulation is in providing fundamental knowledge regarding the function, structure, and connections of the human brain. A common criticism of many studies investigating the neural mechanisms of behavior is that any inferences that are drawn between neural activity and cognitive function are only correlative [33]. Predicting the effects of stimulation can offer the possibility of controlling its effects, and therefore an opportunity to answer this criticism by evoking neural activity in a precise manner and demonstrating the causal role played by such activity on behavior. This approach could complement existing studies of single pulse stimulation and the resulting cortical evoked potentials that have already been used to infer the presence of causal connections between brain regions [17, 18]. Interestingly, how these evoked potentials differ between brain regions has also not been well explored. Future studies could investigate how the features of the response models described here, such as the size, shape, and latency of the linear component, may distinguish the neural architecture and connections between one region and another [25, 26].

Our goal here was to determine whether, by capturing evoked responses to individual stimulation pulses with different amplitudes, we could build a simple model to predict future responses to novel inputs. This approach was premised on a starting assumption that the brain and the response to stimulation can be reasonably approximated as a linear system with responses that are consistent, reproducible, scalable, and that can be combined. The responses that we observe to repeated series of individual pulses suggest that the effects of stimulation are relatively time invariant. Because the dynamics of brain activity are variable and therefore hard to predict, observing consistent responses to stimulation across time is a necessary requirement in order to predict any future response to direct cortical stimulation. Moreover, the evoked activity that our simple model could predict in response to single pulses suggest that the response amplitudes scale linearly with stimulation. If the effects of stimulation are scalable and can be linearly combined, then the responses to multiple individual pulses should reflect the superposition of the responses to individual pulses. We found that using the impulse response to predict the response to multiple pulses can significantly reduce the error in predicting the observed data, providing additional evidence that the assumption of linearity is a reasonable first approximation.

It is clear, however, from the observed evoked responses that the responses to stimulation have a component that is nonlinear. At the very minimum, the response to positive and negative stimulation have similar directions, suggesting an initial rectification of the stimulation input. Such a rectification suggests that perhaps only one phase of the stimulation pulse has modulatory effects on the underlying neural structures. In addition, we find that the derived predictions do not perfectly capture all of the deflections observed in the actual responses to novel inputs. This may be related to nonlinear effects of stimulation that are not captured in our analyses, or that may result from more temporally overlapping responses to higher frequency stimulation. Moreover, we specifically focused on the large evoked response observed during the first 300 ms following stimulation. These responses are dominated by a low frequency component that largely returns to baseline by that time. Although our data suggest that these responses can be reasonably captured using a linear approach, they do not address the extent to which higher frequency smaller amplitude signals that may also be present in the evoked response can be characterized and predicted.

Our data also do not take into account the nonlinearities that may arise when stimulating at multiple locations simultaneously. More sophisticated approaches have been invoked to explore such nonlinearities when characterizing the neural responses to visual inputs [34]. Examining the responses to multielectrode stimulation would be a natural extension of this work, and similar approaches as those use here could help elucidate higher order spatial interactions related to electrical stimulation. In addition, because of stimulation artifacts, we are unable to characterize the response at the actual site of stimulation. This may be addressed through better recording amplifiers that are capable of capturing evoked responses even in the same electrode used for stimulation.

In our analysis, we used the average evoked response at each electrode to identify whether that electrode is responsive to stimulation. Our criteria for determining responsiveness was that the average evoked response should have the characteristic biphasic morphology observed in previous studies of cortico-cortical evoked potentials [16–18], and that the energy of the average evoked response exceeds a given threshold. As expected, predictions were better for those electrodes that were identified as responsive. However, the distinction between responsive and non-responsive electrodes is somewhat arbitrary, and the threshold we choose could fall at any point along a spectrum. The choice of threshold reflects the tolerance for including or excluding electrodes based on a different signal to noise ratio in their response to stimulation. In practice, stimulation at one site likely results in effects at many more electrodes than those we have identified as responsive. Hence, a more general approach for predicting responses that is agnostic to whether an individual electrode is responsive would likely exhibit variable performance in predicting the evoked responses. Reducing the threshold for inclusion may identify a larger set of responsive electrodes, but the reduced signal to noise ratio of their responses may compromise the predictive ability of the response models associated with those newly included electrodes.

Finally, our results suggest that the predictions derived using a simple linear model match the predictions one could make by using the average response to a repeated series of stimuli. An important distinction, however, between using the average response to repeated stimulation to predict future responses and using the linear model is that the latter does not require multiple stimulation sessions to accurately predict future responses. From a practical standpoint, our data therefore suggest that a single and relatively short training session is sufficient for generating a linear model that is capable of predicting the future response to novel stimulation inputs, therefore providing a principled approach for designing and anticipating the responses to different stimulation paradigms.

Conclusion

Our analyses here focus on capturing the responses to stimulation in multiple brain regions when stimulating at an individual site. Our results suggest that the neural responses to direct electrical stimulation are approximately linear and consistent across time, and can therefore be predictable. Characterizing the responses to stimulation in this manner may therefore may provide an important tool for advancing our ability to directly stimulate the human brain in a principled manner.

Declaration of Interests

The authors declare no competing financial interests.

Acknowledgements:

We are indebted to all patients who have selflessly volunteered their time to participate in this study. The views, opinions, and/or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Funding:

This work was supported by the Intramural Research Program of the National Institute for Neurological Disorders and Stroke and partially supported by the DARPA Restoring Active Memory (RAM) program (Cooperative Agreement N66001-14-2-4032).

Author Contributions:

Conceptualization KAZ; Methodology CS, PS, TCS, JHW, KAZ; Software CS, TCS; Validation CS, PS, TCS, SS; Formal analysis: CS, PS, TCS; Investigation CS, TCS, JHW, SKI, KAZ; Resources KAZ; Data Curation TCS, JHW, SKI; Writing - Original Draft CS, KAZ; Writing - Reviewing and Editing CS, PS, JHW, SKI, SS, KAZ; Visualization CS, KAZ; Supervision SS, KAZ; Project Administration KAZ; Funding KAZ.

References

[1] Cohen M, Newsome W (2004) What electrical microstimulation has revealed about the neural basis of cognition. *Current Opinion in Neurobiology* 14(2):169–177.

[2] Borchers S, Himmelback M, Logothetis NK, Karanath H (2012) Direct electrical stimulation of the human cortex - the gold standard for mapping brain functions? *Nature Reviews Neuroscience* 13(1):63–70.

[3] Fisher R, Velsco A (2014) Electrical brain stimulation for epilepsy. *Nature Reviews Neurology* 10(5):261–270.

[4] Nowak L, Bullier J (1998) Axons, but not cell bodies, are activated by electrical stimulation in cortical gray matter. *Experimental Brain Research* 118:477–488.

[5] Rattay F (1999) The basic mechanism for the electrical stimulation of the nervous system. *Neuroscience* 89:355–356.

[6] Tehovnik E, Tolias A, Sultan F, Slocum W, Logothetis NK (2006) Direct and indirect activation of cortical neurons by electrical microstimulation. *J Neurophysiol* 96:512–521.

[7] Histed M, Bonin V, Reid C (2009) Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. *Neuron* 63:508–522.

[8] Logothetis N, et al. (2010) The effects of electrical microstimulation on cortical signal propagation. *Nature Neuroscience* 13(10):1283–1291.

[9] Penfield W, Boldrey E (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60(4):389–443.

[10] Brindley GS, Lewin W (1968) The sensations produced by electrical stimulation of the visual cortex. *Journal of Physiology* 196:479–493.

[11] Sanai N, Mirzadeh Z, Berger M (2008) Functional outcome after language mapping for glioma resection. *New England Journal of Medicine* 358:18–27.

[12] Suthana N, et al. (2012) Memory enhancement and deep-brain stimulation of the entorhinal area. *The New England Journal of Medicine* 366:502–510.

[13] Jacobs J, et al. (2016) Direct Electrical Stimulation of the Human Entorhinal Region and Hippocampus Impairs Memory. *Neuron* 92(5):983–990.

[14] Ezzyat Y, et al. (2017) Direct Brain Stimulation Modulates Encoding States and Memory Performance in Humans. *Current biology: CB* 27(9):1251–1258.

[15] Desmyrget M, et al. (2009) Movement intention after parietal cortex stimulation in humans. *Science* 324(5928):811–813.

[16] Matsumoto R, et al. (2004) Functional connectivity in the human language system: a cortico-cortical evoked potential study. *Brain* 27:2316–2330.

[17] Kunieda T, Yamao Y, Kikuchi T, Matsumoto R (2015) New approach for exploring cerebral functional connectivity: Review of cortico-cortical evoked potential. *Neurol Med Chir* 55:374–382.

[18] Keller C, et al. (2014) Mapping human brain networks with cortico-cortical evoked potentials. *Philosophical Transactions of the Royal Society B* 369(20130528).

[19] Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29:162–173.

[20] Trotta MS, et al. (2017) Surface based electrode localization and standardized regions of interest for intracranial eeg. *Human brain mapping* 39:709–721.

[21] Marmarelis P, Marmarelis VZ (1978) *Analysis of Physiological Systems.* (Springer, Boston MA).

[22] Ljung L (1987) *System Identification: Theory for the User.* (P T R Prentice Hall).

[23] Benabid A, et al. (2004) Deep brain stimulation for movement disorders in *Youman's Neurosurgical Surgery, 5th Edition*, ed. Winn H. (Saunders, Philadelphia) Vol. 3, pp. 2803–2827.

[24] Kocobicak E, Temel Y, Hollig A, Falkenburger B, Tan S (2015) Current perspectives on deep brain stimulation for severe neurological and psychiatric disorders. *Neuropsychiatric Disease and Treatment* 11:1051–1066.

[25] Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10(3):186–198.

[26] Chapeton JI, Inati SK, Zaghloul KA (2017) Stable functional networks exhibit consistent timing in the human brain. *Brain* 140(3):628–640.

[27] Tehovnik E (1996) Electrical stimulation of neural tissue to evoke behavioral responses. *Journal of Neuroscience Methods* 65:1–17.

[28] Santaniello S, et al. (2014) Therapeutic mechanisms of high-frequency stimulation in parkinson's disease and neural restoration via loop-based reinforcement. *Proc Natl Acad Sci U S A* 112(6):E586–95.

[29] Nakae T, et al. (2018) Oscillatory responses evoked by single-pulse electrical stimulation in human cerebral cortex. *Clinical Neurophysiology* 129(189-190).

[30] Victor JD (1987) The dynamics of the cat retinal x cell centre. *Journal of Physiology* 386:219–246.

[31] Chichilnisky E (2001) A simple white noise analysis of neuronal light responses. *Network* 12:199–213.

[32] Freeman D, Rizzo III J, Fried S (2010) Electric stimulation with sinusoids and white noise for neural prostheses. *Frontiers in Neuroscience* 4:1.

[33] Krakauer J, Ghazanfar A, Gomez-Marin A, MacIver M, Poeppel D (2017) Neuroscience needs behavior: Correcting a reductionist bias. *Neuron* 93(3):480–490.

[34] Sutter E (2001) Imaging visual function with the multifocal m-sequence technique. *Vision Research* 41:1241–1255.

Figure Legends

Figure 1 Evoked Responses to Single Pulse Stimulation. A) Example stimulation pulse and evoked response recorded in a single electrode. **B)** In an example participant, stimulation was delivered at one bipolar cortical site (*white*). Only a subset of electrode (*green*) are responsive. **C)** Stimulation consisted of stimulation pulses with amplitudes drawn from a white noise distribution. A single electrode exhibits evoked responses to the pulses that appear modulated by stimulation amplitude. **D)** For a single responsive electrode, responses to all stimulation amplitudes over time showed consistently timed, symmetric response to high amplitude current stimulation. **E)** For this same electrode, the energy of the evoked response was also symmetric with stimulation amplitude and increases with greater magnitude of stimulation. **F)** Stimulation pulses were delivered to each participant for five minutes of training at a single cortical location. During testing, a series of pulses with novel amplitudes was delivered for two minutes at the same location. The test series of pulses was repeated 3-5 times in each participant.

Figure 2 Responses to Repeated Stimulation. A) We repeated the two-minute series of testing pulses three to five times in each participant. The responses at a typical recording electrode were consistent. Using all repetitions excluding the first one, we averaged the responses to create a predicted response based on the average. We compared this average response to the actual response observed during the first repetition. **B)** The MSE between the average response and the first repetition (Average Response Residual) was significantly smaller than the mean of squares of the signal, which is equivalent to the error if our predicted response were the mean value of zero ($\mathrm{MS}_{\mathrm{signal}}$, Signal), but was higher than the variance of the baseline noise activity ($\mathrm{MS}_{\mathrm{noise}}$, Noise). **, $p < .01$; *, $p < .05$; paired t-test.

Figure 3 Linear System Model Predicts the Responses to Single Pulses. A) When constructing the simple model characterizing the response to stimulation, we divided each training series of pulses, $x[t]$, into positive and negative pulses because the responses to stimulation appeared symmetric. We therefore modeled the system as a nonlinear rectification followed by a separate impulse response, $h^{+}[t]$ and $h^{-}[t]$ for positive and negative pulses, that together produce the output $y[t]$. **B)** Cross-correlating the input $x[t]$ of stimulation pulses with the recorded output, $y[t]$, generates the first order kernel, or impulse response $h[t]$, of the system for each recorded electrode. **C)** For each recorded electrode, we can convolve any new input $x[t]$ with the derived model to generate a prediction, $\hat{y}[t]$ of the evoked response. The model prediction appears to match the actual evoked response, $y[t]$, to the novel series of pulses. **D)** Across responsive electrodes, the MSE between the predicted and recorded response (Model Error) was significantly lower than the mean of squares of the signal, which is equivalent to the MSE if the predicted response were the mean value of zero, ($\mathrm{MS}_{\mathrm{signal}}$, Signal). The residual error was higher than the mean of squares of the baseline activity, which is the lowest possible level of MSE that is attainable ($\mathrm{MS}_{\mathrm{noise}}$, Noise). The model prediction reduced the error from its maximum value to its minimum possible value by $69.7 \pm 12.1\%$ across responsive electrodes.

, $p < .01$; paired t-test. **E) The MSE for each responsive electrode using the model prediction was similar to the MSE derived using the prediction based on the average response. Each point represents a single responsive electrode.

Figure 4 Predicting Responses to Multiple Pulses. A) In a subset of participants, each trial during testing contained between two to seven pulses in quick succession (*grey*; three example trials shown). Using the derived model, we predicted the evoked response in a single electrode to the multiple pulses in different trials. The correspondence between the predicted response (*blue*) and the recorded response (*black*) suggests that predicting the response to multiple pulses can be attained through superposition. **B)** Across responsive electrodes, the MSE between the model prediction and the recorded responses to multiple pulses (Model Error) was significantly lower than the mean of squares of the signal, which is equivalent to the MSE if the predicted responses were the mean value of zero, ($\mathrm{MS}_{\mathrm{signal}}$, Signal). The model prediction reduced the error for predicting multiple pulses from its maximum value to its minimum possible value ($\mathrm{MS}_{\mathrm{noise}}$, Noise) by 88.6 ± 22.9% across responsive electrodes. **, $p < .01$; *, $p < .05$; paired t-test. **C)** The MSE for each responsive electrode using the model prediction of the response to multiple pulses was similar to the MSE using the prediction based on the average response. Each point represents a single responsive electrode.

Supplementary Material and Methods

Experimental Design

While participants were awake, at rest, and monitored in the Epilepsy Monitoring Unit, we applied single pulses of electrical stimulation to a single pair of subdural electrode contacts. In two individuals, we performed the same experiments in two separate stimulation locations. We considered the data from each separate location as a separate and independent dataset. We therefore report the results of 12 unique stimulation sites used for analysis.

This experimental setup is similar to previous studies of cortico-cortical evoked potentials [16, 18]. We used a programmable neurostimulator (CereStim, Blackrock Microsystems, LLC., Salt Lake City, UT) and a custom built GUI to deliver individual biphasic stimulation pulses through selected subdural contacts. In this case, each pulse consists of a square-wave biphasic pulse, where each phase has a duration of 0.3 ms (Fig. 1A). The two phases of each pulse were separated by a gap of 0.05 ms. We used a biphasic pulse to avoid electrical charge buildup on the cortical surface over the course of the stimulation sessions (for safety) and to avoid polarization of the electrode contacts which could reduce current density [2]. The amplitudes of both phases of each biphasic pulse were equal, and we defined a positive (negative) amplitude as one in which the biphasic pulse started with a positive (negative) deflection.

In order to characterize the responses to electrical stimulation, we applied individual biphasic pulses to a pair of neighboring electrodes while recording the intracranial EEG (iEEG) responses in the remaining electrodes in each participant. In most participants, we delivered these pulses once every 800-1000 ms with a random jitter of 10 ms, resulting in a stimulation frequency of approximately 1 Hz. In two participants, we used an inter-stimulus interval of 630 ms. The average inter-stimulus interval across all participants was 867 ± 41.3 ms. We used these approximately one second intervals between stimulation pulses so that the typical evoked response following each stimulation pulse had sufficient time to return back to baseline. We considered the presentation of each individual stimulation pulse, followed by its inter-stimulus interval, a single stimulation trial. For each individual pulse, we used a stimulation amplitude that was randomly drawn from a uniform distribution of amplitudes between approximately 8 and -8 mA (Fig. 1C). Each series of trials was therefore comprised of pulses with stimulation amplitudes that are temporally uncorrelated from one another and zero-mean, and that can be described as a white noise input.

We presented trials of single pulses of electrical stimulation in two experimental sessions while capturing iEEG recordings in each participant (Fig. 1F). Given the approximately one second interval between trials, in each session the stimulation frequency was approximately 1 Hz. In the first session, which we refer to as training, we presented a series of trials of single pulses for five minutes, producing evoked responses to approximately 300 individual stimulation pulses with amplitudes drawn from the uniform distribution. We used this training session in order to generate a tractable model that characterizes how each recording site responds to stimulation at each stimulation site (see Linear Systems Model of the Evoked Response). In the second session, which we refer to as testing, we presented novel trials of individual pulses, consisting of novel stimulation amplitudes also drawn randomly from the same distribution, for two minutes (approximately 120 individual stimulation pulses). We used these novel test pulses and their responses to test how well the identified model can predict the evoked response to novel stimulation amplitudes. We repeated the same test series of trials of pulses three to five times (3.67 ± .28 repetitions per participant) in order to assess whether the evoked responses to the individual pulses were preserved across repetitions.

In five participants, during testing we also delivered a two-minute series of trials containing multiple pulses rather than only individual pulses. Our goal was to determine whether our simple linear model could predict the compound response to multiple pulses. In these testing sessions, some stimulation trials consisted of between two and seven pulses in close succession (Fig. 4A), while the remaining trials involved only a single pulse of stimulation. For the trials with multiple pulses, each pulse was presented within 20 to 50 ms of the previous pulse, and the final pulse was followed by the same inter-stimulus interval that we used when presenting single pulses during the training session (approximately 867 ms on average). Hence, while the trials had an overall stimulation frequency of approximately 1 Hz given the inter-stimulus interval between trials, within the trials with multiple pulses the frequency of stimulation was not fixed. We used the trials with multiple pulses to test how well the derived model can predict the evoked response to multiple pulses. As above, we repeated this presentation of multiple pulses between three and five times in a subset of participants (5 participants, 3 ± .32 repetitions per participant). We also extracted the trials with single pulses in these testing sessions to contribute to our dataset examining how well the derived model can predict the evoked response to single pulses. In total, therefore, we collected data from five participants in whom we tested the predictions in response to multiple pulses, but by extracting trials involving only a single pulse, we were able to analyze the predictions to single individual pulses in all twelve stimulation sites.

Intracranial Recordings and Pre-Processing

We applied single pulse electrical stimulation at each stimulation site, which consisted of two adjacent electrode contacts, in a bipolar fashion so that electrical current flow was localized to areas of cortex below the electrode contacts. We chose electrode locations for stimulation in each participant based on two criteria. First, we confirmed with the clinical team that electrode locations were not directly involved in seizure activity. Second, we chose electrode locations for stimulation that were relatively central to the entire set of implanted electrodes in a given participant so as to maximize the chances of observing evoked responses at other electrode contacts.

We recorded iEEG data in response to the input pulses of electrical stimulation from subdural contacts (PMT Corporation, Chanhassen, MN) sampled at 1000 Hz. Subdural contacts were arranged in both grid and strip configurations with an inter-contact spacing of 10 mm. In all cases, placement of the contacts was determined by the clinical team in order to best localize epileptogenic regions for resection. Across all participants, we captured iEEG from a total of 1152 electrode contacts. We arranged electrodes on the cortical surface based on the location of the hypothesized epileptic focus. We localized each contact by co-registering the postoperative CTs with the postoperative MRIs using the publicly available software package AFNI [19]. We subsequently projected the resulting contact locations to the cortical surface of a Montreal Neurological Institute standard brain to identify anatomic locations (Fig. 1B and Supplementary Fig. S2) [20].

Electrical stimulation results in a stimulation artifact in the recorded trace that is present at the beginning of each trial. In order to determine the time points that were contaminated by this stimulation artifact, we examined the mean evoked response in each electrode, averaged across all stimulation trials, and identified the time point of the average initial sharp deflection observed following delivery of the stimulation pulse. We defined the stimulation artifact window as the first 10 ms following stimulation and discarded these time points from the recorded trace in subsequent analyses so as to avoid contamination from any stimulation artifacts.

For each electrode, we applied a high pass filter at 2 Hz to remove noise due to electrode drift, a 60 Hz notch filter to remove line noise, and a low pass filter at 200 Hz to remove higher frequency noise. In each experimental session, we identified any recording electrode with a voltage trace whose variance was more than one standard deviation away from the median variance of all recording electrodes and excluded those electrodes from further analysis. Removing these trials ensures that any trials with obvious amplifier saturation are excluded from our analysis. We also removed any electrodes with recording traces that appeared obviously flat on visual inspection. We did not note any afterdischarges with stimulation during any experimental session, and so did not exclude any electrodes that were identified as being involved in seizure activity from our analysis. We thus retained 1004 electrodes for further analysis. In each participant, we re-referenced the raw recorded trace of each electrode by subtracting the common average signal averaged across all of the retained electrodes. We divided the continuous iEEG recording for each session into individual response trials for each recording electrode, where each trial captures the evoked response to one single stimulation pulse. Hence, each trial represents approximately 867 ms of iEEG data, depending upon the inter-stimulus interval for that participant.

Identifying Responsive Electrodes

We captured the evoked response to every stimulation pulse during the training session. We examined the mean evoked response, $y[t]$, in̄ each recorded electrode over all stimulation trials to determine if that electrode had a meaningful response to stimulation for a majority of stimulation amplitudes. Examining the average response across all trials implicitly gives more weight to the trials with high stimulation amplitudes, since their evoked responses have greater signal to noise. We had two requirements for an electrode to be considered responsive.

First, we required that the average evoked response have a minimum of two time points during the first 300 ms during which the value of the average evoked response crosses zero. This ensures that the average evoked response has the typical bi- or tri-phasic morphology observed in previous studies of cortico-cortical evoked potentials [16–18]. This requirement also ensures that the recordings from any electrodes that exhibit obvious amplifier saturation, which would result in a slowly decaying signal back to baseline, are excluded since they would not exhibit the minimum of two time points during which the response crosses zero.

Second, we required that the ratio of the energy contained within the first 300 ms (after the 10 ms artifact) of the average evoked response, $\sum_{11}^{310} \overline{y}[t]^2$, compared to the energy contained in the next 300 ms of the average evoked response, $\sum_{311}^{610} \overline{y}[t]^2$, exceed a specific threshold (Supplementary Fig. S1). Most evoked responses that we observed occurred within 300 ms of stimulation, and so we considered the activity of each electrode recorded beyond 300 ms in each trial to have returned to baseline, consistent with the time course of evoked responses observed in previous studies of cortico-cortical evoked potentials [16–18]. To confirm that this was the case, we examined the energy during this time period beyond 300 ms in all trials with stimulation amplitudes less than 1 mA. These trials do not evoke an appreciable response in the recorded traces, and so we take this epoch in these trials to reflect the natural fluctuations of the signal at rest. We compared the energy in these epochs in these low amplitude trials to the same epochs in the remaining trials, and found that they were similar ($0.024 \pm 0.006\text{mV}^2$ versus $0.027 \pm 0.009\text{mV}^2$), suggesting that activity during the second 300 ms window in all trials indeed reflects the baseline activity. This requirement therefore ensures that the average evoked response of that electrode is comprised of sufficient signal relative to the baseline level of activity. In order to determine this threshold, we computed this ratio for every electrode in every participant. We rank ordered the electrodes by the ratio of energy between the two time periods (Supplementary Fig. S1C). The distribution of ratios had a clear knee, above which the electrodes had a larger ratio of energy. This knee suggested a natural threshold for retaining electrodes with sufficient energy. To calculate this threshold, we fit two lines to the two regions of the distribution and designated the ratio at which the lines crossed as the threshold. This threshold corresponded to an energy ratio of 5.85. All electrodes with average evoked responses that had a minimum of two zero-crossings and a ratio of energy that exceeded this threshold were retained for further analysis as responsive electrodes.

In order to examine whether the evoked responses were symmetric, we characterized the energy of the evoked response, on average, to positive and negative stimulation over the first 300 ms of each trial, $E^+ = \sum \overline{y}^+[t]^2$ and $E^- = \sum \overline{y}^-[t]^2$. We defined an asymmetry index for each responsive electrode based on these energies: $AI = \frac{E^+ - E^-}{E^+ + E^-}$. AI will be greater than zero if an electrode has a stronger energy of the impulse response to positive stimulation compared to negative stimulation, and will be approximately zero if the responses are similar to both stimulation polarities. If an electrode only has a response to positive (negative) stimulation, the AI will be close to 1 (-1).

Linear System Model of the Evoked Response

For each responsive electrode, we constructed a model that characterizes the evoked response to stimulation at the stimulation site. Each model had an identical architecture, comprised of a linear impulse response that separately operates on positive and negative input pulses. We generated each model by examining the evoked responses to the input pulses presented during the five-minute training session in each participant.

The input to our model is comprised of a static nonlinear gain that rectifies and therefore selects for positive and negative stimulation inputs, respectively (Fig. 3A). We separately rectified the positive and negative stimulation pulses as inputs to the linear model because of the observation that the evoked responses to positive and negative stimulation were often similarly shaped with deflections in a similar direction (Fig. 1C,D). Although some electrodes had asymmetric response magnitudes depending on the polarity of the stimulation input (Supplementary Fig. S2C), we set the gain of this nonlinear rectification to one for positive inputs and to negative one for negative inputs for simplicity. Any differences in the magnitude of the gain of the response to positive and negative stimulation should be captured by the linear operators in our model.

Given the similarly shaped evoked responses observed following each stimulation pulse and the relatively consistent responses we observed following the repeated trials of stimulation pulses, we used a linear time-invariant (LTI) operator to characterize the response to stimulation (Fig. 3A). This component of the model assumes that the responses to stimulation, following rectification, scale linearly with stimulation amplitude. Our analysis examining how well the responses predicted by this model during testing match the recorded traces was directly designed to investigate this assumption. Under this assumption, we can derive the first order kernel of the system transfer function, or the impulse response $h[t]$, that fully characterizes the response at every recording site to stimulation by capturing the evoked responses to the input pulses presented during training (Fig. 3B). Because of the rectification of stimulation input, we separately derived the impulse responses for inputs of positive and negative polarity, $h^+[t]$ and $h^-[t]$, respectively.

Hence, given an input, $x[t]$, our model estimates a predicted output, $\hat{y}[t] = \mathcal{H}\{x[t]\}$ by separately convolving the rectified positive and negative input pulses with the impulse responses, $h^+[t]$ and $h^-[t]$, respectively:

$$\hat{y}[t] = \mathcal{H}\{x[t]\} = \sum_{\tau=0}^{L} h^+[\tau] f^+(x[t-\tau]) + \sum_{\tau=0}^{L} h^-[\tau] f^-(x[t-\tau]),$$

where $f^+(x)$ and $f^-(x)$ are the rectifiers that select positive and negative pulses, respectively, and where each impulse response is of finite duration, L. The model $\mathcal{H}$ is therefore fully parametrized by the two finite impulse responses, $h^+[t]$ and $h^-[t]$.

To estimate the impulse responses, $h^+[t]$ and $h^-[t]$, that provides the best model for each electrode, we first idealized the input stimulation pulses such that each pulse on each trial, k, had amplitude a_k, and occurred T ms following the previous pulse:

$$x[t] = \sum_{k} a_k \, \delta[t - kT] \quad \text{with } a_k \sim U(-8, 8).$$

We defined this input series of amplitudes as our input $x[t]$ at the chosen stimulation site. The amplitudes of each pulse, a_k, were drawn from a uniform distribution of amplitudes between -8 and 8 mA. This input is therefore a series of stimulation amplitudes that are uncorrelated random variables and have zero mean and finite variance. In the idealized input series, each pulse occurs every T ms, representing the intertrial interval, although in practice each pulse followed the previous pulse by $T \pm 10$ ms, reflecting the random jitter on each trial. During the training experimental session, the input pulses, $x[t]$, resulted in a series of observed evoked responses, $y[t]$, that occurred every T ms and that we recorded in every electrode. Our goal was to estimate a model, $\mathcal{H}$, for each responsive electrode based on observing the response $y[t]$ of our system to this input $x[t]$.

A widely used approach for identifying the best model for each responsive electrode is to estimate the model that minimizes the mean-square error between the predicted response $\hat{y}[t]$ and the observed response $y[t]$. The mean-square error of the predicted response generated by the model $\mathcal{H}$ is defined as:

$$\epsilon(\mathcal{H}) \triangleq E(y[t] - \hat{y}[t])^2$$

where $E(..)$ denotes expected value. The problem of finding the minimum mean-square estimate of the model can then therefore be expressed as one of minimizing ϵ. For each LTI operator, this problem is equivalent to estimating the impulse responses, $h^+[t]$ and $h^-[t]$.

The optimal solution for the impulse responses that minimize this mean-squared error can be estimated by simply cross-correlating the input and output of each LTI operator and normalizing by the autocorrelation of the input at lag $\tau = 0$ [21, 22]. For a series of input pulses, the autocorrelation at lag $\tau = 0$ is equal to the energy of the signal and is simply the sum of the square of all of the amplitudes, a_k, used in the series. We can therefore estimate the impulse response, $h[t]$, for the LTI operator that minimizes the mean-square error simply by cross-correlating the rectified input series of pulses with the recorded output and normalizing by the energy of the input (Fig. 3B).

Because the duration of each finite impulse response is smaller than the duration of a single trial ($L < T$), and because the input in each trial during training is either positive or negative, we were able to decouple the minimization over $h^+[t]$ from the minimization over $h^-[t]$ and estimate each impulse response independently. To do so, we divided the continuous time series of stimulation inputs, $x[t]$, and evoked responses, $y[t]$, into individual trials, where each trial captures the evoked response to one single stimulation pulse (approximately 867 ms in length on average), and separated the trials into two groups based on the polarity of the stimulation amplitude (positive or negative stimulation). We concatenated all of the trials in each group to create two separate input series of stimulation amplitudes, $x^+[t]$ and $x^-[t]$, and two separate continuous output streams of evoked responses, $y^+t]$ and $y^-[t]$. We used these separate inputs and outputs to estimate $h^+[t]$ and $h^-[t]$, the impulse responses to positive and negative stimulation, respectively (Fig. 3A).

Statistical Analyses and Prediction of Evoked Responses

Using the derived impulses responses, $h^+[t]$ and $h^-[t]$, we then estimated a predicted output, $y\hat{[t]}$, to the input pulses presented during the testing experimental session. We computed the prediction error between $\hat{y}[t]$ and the observed output $y[t]$ by calculating the mean squared error (MSE) directly on the differences between their time series. The MSE of any difference between the predicted and recorded output in each trial is given by:

$$\mathrm{MSE} = \frac{1}{N} \sum_{t=11}^{300} (y[t] - \hat{y}[t])^2$$

where $N = 290$ is the length of the window over which we compute the error in each trial. Of note, for each of the individual trials that together constitute the continuous traces $\hat{y}[t]$ and $y[t]$, the model provides meaningful information only over the duration of the evoked response, which typically is less than 300 ms. Moreover, any predictions cannot account for the stimulation artifact that occurs immediately after stimulation. For the predictions of single pulses, we therefore only used a time window $11 \leq t \leq 300$ ms in every trial for calculating the prediction error. This time window in each trial excludes the stimulation artifact observed in the first 10 ms and any time points greater than 300 ms following stimulation from our measure of error. For inputs comprised of multiple pulses, however, different trials contained a different number of pulses. In these cases, we were primarily interested in the response to stimulation following completion of all of the pulses in each trial, and therefore used a time window beginning 10 ms after the final pulse in each trial, to exclude the stimulation artifacts, up to 300 ms following that final pulse. We also examined how well the model could predict the response to stimulation in the intervals between the pulses, although in this case we again excluded the first 10 ms stimulation artifact after each pulse (Supplementary Fig. S5B). In all cases, we calculated this error for every trial during the testing session, and then computed the average across trials to generate a MSE for each recording electrode.

We compared the MSE between the predicted and recorded trace to mean of squares of the evoked response, $y[t]$, which is proportional to the energy of the observed response. We refer to this as $\mathrm{MS}_{\mathrm{signal}}$:

$$\mathrm{MS}_{\mathrm{signal}} = \frac{1}{N} \sum_{t=11}^{300} y[t]^2$$

We use this as the upper bound of MSE since it is equivalent to the error we would expect had we made no prediction, or if our predicted trace were the mean value of the recorded trace, which was zero after pre-processing the raw data, for all time. To compute this upper bound and to compare it to the MSE error observed with our model prediction, we used the same time window on each trial, which is the first 300 ms of time series data, excluding from this time window the time points up to 10 ms corresponding to the stimulation artifact time ($11 \leq t \leq 300$ ms). As with the model predictions, we calculated this upper bound of error for every trial during the testing session, and then computed the average across trials to generate a maximum error $\mathrm{MS}_{\mathrm{signal}}$ for each recording electrode.

We also compared the MSE between the predicted and recorded trace to the mean of the squares of the baseline activity for each electrode, which is proportional to that electrode's baseline energy. We refer to this as $\mathrm{MS}_{\mathrm{noise}}$, and this represents the lower bound of MSE one could achieve given any baseline level of noise in the recorded trace. To compute this lower limit of error, we computed the mean of squares of the recorded output, $y[t]$, using the final 300 ms of time series data from each trial, and again averaged across all trials. This time period corresponds to the end of each trial, by which time any evoked responses would have returned to baseline activity. In practice, in order to fairly compare the MSE of the predicted response to this lower limit of error, we used the final 290 ms to match the same number of time points used in the evaluation of the prediction error for each electrode.

Using both the upper and lower bounds for MSE, for each responsive electrode we then computed how much the model prediction reduces the error from the upper bound $\mathrm{MS}_{\mathrm{signal}}$ to the lower bound $\mathrm{MS}_{\mathrm{noise}}$ by expressing that reduction as a percentage:

$$(1 - \frac{\mathrm{MSE} - \mathrm{MS}_{\mathrm{noise}}}{\mathrm{MS}_{\mathrm{signal}} - \mathrm{MS}_{\mathrm{noise}}}) \times 100$$

where MSE is the measure of error between the predicted trace, $\hat{y}[t]$, and observed trace, $y[t]$, for that electrode. Given the baseline level of noise, the best prediction achievable would reduce the error from the maximum level, $\mathrm{MS}_{\mathrm{signal}}$, proportional to the energy of the evoked response, to the minimum level, $\mathrm{MS}_{\mathrm{noise}}$, proportional to the energy of the noise, by 100% in each electrode.

Supplementary Figure Legends

Figure S1 Identifying Responsive Electrodes. A We identified responsive electrodes using two criteria. A typical responsive electrode, like the one depicted, needs to exhibit an average evoked response with more than two zero crossings (green dots) and signal energy that was higher in the first 300 ms after response (yellow) compared to the energy during the next 300 ms (blue), excluding the artifact zone (blue dashed line). **B** A typical unresponsive electrode, as depicted, may have an average evoked response that exhibits some energy and zero crossing points. **C** We rank ordered all of the electrodes by ratio of signal energy between the first and second 300 ms time points. We determined an energy ratio threshold of 5.85. Electrodes with energy ratios greater than this threshold were considered responsive. We eliminated electrodes with too few zero crossings regardless of their energy ratio.

Figure S2 Responsive Electrodes across Participants A Responsive electrodes (green) and the stimulation sites (white) were identified on surface reconstructions of each participant's brain. **B** Across participants, responsive electrodes were mostly local. However, some distal electrodes were also responsive. **C** Most responsive electrodes were shown to be equally responsive to positive and negative pulses, but a subset

was more responsive to negative stimulation and therefore had a negative asymmetry index.

Figure S3 Responses to Repeated Trials of Single Pulse Stimulation. A In a typical example of a responsive electrode, the evoked response to stimulation pulses were similar when repeating the trials of pulses several times ($y_1[t]$ to $y_4[t]$). The consistency in the responses suggests that the response to stimu-lation is relatively time-invariant. **B** A single trial from the series of pulses demonstrates the consistency of the response across two repetitions. **C** When examining all 1004 electrodes (including both responsive and non-responsive electrodes), the MSE when using a prediction based on the average response (Average Response Residual) is significantly lower than the mean of squares of the recorded response, which represents the upper bound of error (MS$_{signal}$, Signal; 0.118 ± 0.044 mV^2 versus 0.194 ± 0.055 mV^2, $t(1003) = -3.16$, $p = .002$, paired t-test). The prediction is not significantly different than the mean of squares of the baseline activity, which is the lower bound of error activity (MS$_{noise}$, Noise; 0.118 ± 0.044 mV^2 versus 0.040 ± 0.007 mV^2, $t(1003) = 1.80$, $p = .072$, paired t-test). This indicates a consistency in response to electrical stimula-tion over time. **, $p < .01$; paired t-test.

Figure S4 Prediction Error for Responsive and Non-responsive Electrodes for Single Pulses. A In responsive electrodes ($n = 52$), the linear model prediction and the prediction based on the average response showed no significant difference in MSE error when predicting the response to single pulses. The differences between the model prediction and upper and lower limits of error, MS$_{\text{signal}}$ (Signal) and MS$_{\text{noise}}$ (Noise) respectively, were significant. For non-responsive electrodes ($n = 952$), the model prediction and the prediction based on the average response showed no significant difference in MSE, while differences from the MS$_{\text{signal}}$ and MS$_{\text{noise}}$ were statistically significant. **B** The MSE between the predicted response and the observed response to trials of single pulse stimulation is consistent across all trials with different stimulation amplitudes. **C** Using only a subset of trials during the training experimental session can still be sufficient to generate a linear model that can predict the evoked response during testing with MSE values that are similar to the predictions that are generated when using all of the training trials (300 trials). **D** When using a quadratic input that rather than a simple rectification, the extent to which the model can predict the evoked response is similar. The differences between the model prediction and upper and lower limits of error, MS$_{\text{signal}}$ (Signal) and MS$_{\text{noise}}$ (Noise) respectively, were significant. The model prediction and the prediction based on the average response showed no significant difference in MSE error when predicting the response to single pulses (0.029 ± 0.006mV^2 versus 0.024 ± 0.003mV^2, $t(51) = 1.21$, $p = .23$, paired t-test. **, $p < .01$; *, $p < .05$.

Figure S5 Prediction Error for Multiple Pulses for Responsive and Non-Responsive Electrodes. A In the primary analysis, we examined the model prediction to the recorded trace following the final pulse that was delivered in trials with multiple pulses (*top*). In responsive electrodes ($n = 22$), the model prediction and the prediction based on the average response showed no significant difference in MSE error

when predicting the response to multiple pulses. The differences between the model prediction and upper and lower limits of error, $\mathrm{MS}_{\mathrm{signal}}$ (Signal) and $\mathrm{MS}_{\mathrm{noise}}$ (Noise) respectively, were significant. For non-responsive electrodes ($n = 437$), the model prediction and the prediction based on the average response showed no significant difference in MSE, while differences from the $\mathrm{MS}_{\mathrm{signal}}$ and $\mathrm{MS}_{\mathrm{noise}}$ were statistically significant. **B** We additionally examined how well the model prediction compared to the recorded trace when we examined the time series that included all of the pulses in each trial with multiple pulses (*top*). For the time points between individual pulses, we removed the 10 ms stimulation artifact. In responsive electrodes ($n = 22$), the model prediction exhibited a MSE that was significantly less than the upper limit of error, $\mathrm{MS}_{\mathrm{signal}}$ (Signal), ($t(21) = -2.34$, $p = .029$, paired t-test), but significantly greater than the mean of squares of the baseline activity, $\mathrm{MS}_{\mathrm{noise}}$ (Noise), which represents the lower limit of error ($t(21) = 3.62$, $p = .0016$). The model prediction also exhibited an MSE that was significantly larger than the prediction based on the average response ($t(21) = 2.96$, $p = .0076$). For non-responsive electrodes ($n = 437$), the model prediction and the prediction based on the average response showed no significant difference in MSE ($t(436) = 1.87$, $p = .062$), while differences from the $\mathrm{MS}_{\mathrm{signal}}$ (Signal) and $\mathrm{MS}_{\mathrm{noise}}$ (Noise) were statistically significant ($t(436) = -2.91$, p $= .0039$ and $t(436) = 2.08$, $p = .039$). **, $p < .01$; *, $p < .05$; paired t-test.