The Volitional (In)significance of Neuroscience: What Libetian Investigations Can and Cannot Do for Free Will

A thesis presented by
Garrett Lam

to
the Faculty of the committee on Degrees in Neurobiology and Philosophy
in partial fulfillment of the requirements
for the degree with honors
of Bachelor of Arts
and certificate in Mind, Brain, & Behavior

Harvard University,
Cambridge, Massachusetts
March, 2016

Acknowledgements:

On the final paper of my freshman seminar class, my professor wrote: "One senses that you are working through a tension between science and humanity," with a nudge that I choose one and "pursue it in depth." And he was right about the tension. I found in science an extraordinary method to discover the utterly bizarre objects in the world and address the problems that needed to be *solved*, the questions that were truly important to *answer*. But I found in philosophy the actual window that I would see the world through, and philosophy (and the humanities) always seemed to touch closer to the questions I thought were truly important to *ask*. If I wanted to make practical leaps in understanding and create something that could do a lot of good, I felt deeply allured to the power of science. If I got my heart broken, I was turning to a poem, not an article in Nature.

I guess, then, it is only fitting that my final substantive paper at Harvard should reaffirm my complete inability to decide between the two. But of course we can arrive where we once started and know a place for the first time, and in so far as the past four years in culminate (they don't) in some hackneyed way (that is true) in this thesis, I have the following to thank:

To Andrew Sanchez, and my blockmates John Lee, John Griffin, and Jerry Chang, for all the laughs that made this that much easier.

To Grace Huckins, without whom neither this thesis nor my senior year would have been possible, yet alone complete.

To Mom, Dad, Brandon, and Austin, for more than could possibly be put here, or in words.

To Selim Berker, for pulling me to philosophy before it was too late. And Susanna Rinard, for spending our thesis meetings on much more important topics than the philosophy of free will.

To Gabriel Kreiman and Hanlin Tang, for far more support than I deserve.

List of Contributions:

The particular microwire signal collection paradigm variation of Libet experiments was by designed Itzhak Fried, Roy Mukamel, and Gabriel Kreiman. After surgical procedures (i.e. depth electrode placement) were conducted, running of experiments and collection of data was done and overseen by Itzhak Fried, with dataset provided by Gabriel Kreiman. **Figure 3B** was generated by Itzhak Fried, Roy Mukamel, and Gabriel Kreiman. All data preprocessing and filtering was completed independently by Garrett Lam. All analyses, including SVM training, of the dataset were completed independently by Garrett Lam.

Abstract:

Volition and self-initiated behavior are critical components of cognitive control, and the extent to which humans possess free will has implications in moral, legal, and clinical settings. While some progress has been made in variations of the Libet paradigm, in which brain activity is temporally compared to subjects' reports of conscious decisions, neural metrics have been historically limited to extracranial metrics and, recently, spiking. Therefore, the relative roles of subthreshold neuromodulatory activity, such as neural oscillations across the various frequency bands characteristic of local field potentials, remain poorly understood. To examine the neural dynamics underlying volition at this level of granularity, we exploited the spatiotemporal resolution of intracranial recordings in patients with pharmacologically intractable epilepsy during Libet task performance. We observed significant differences in spectral activity between the baseline period and the period leading up to the conscious decision, particularly across beta and high gamma frequencies in the frontal lobe, namely the anterior cingulate cortex (ACC) and the supplementary motor area (SMA). We also report the first ever coherence studies during Libet task performance, providing evidence of interregional synchronicity at the beta and high gamma bands. After flagging several conceptual shortcomings of Libet's paradigm, we dissociate the relative roles such experiments could play in the free will debate, and propose that while such experiments could theoretically cast doubt upon some libertarian varieties of free will, even the best neuroscience leaves much of the dialectic untouched.

Table of Contents

Title Page 1

Acknowledgements 2

List of Contributions 3

Abstract 4

Table of Contents 5

List of Figures 6

Abbreviations 7

Introduction 8

Methods 19

Results 25

Discussion: Empirical Reflections 41

Discussion: Can Libet experiments revolutionize the free will debate? 45

References 65

List of Tables and Figures

Figure 1 12

Figure 2 17

Table 1 17

Figure 3 20

Figure 4 27

Table 2 29

Figure 5 31

Figure 6 32

Figure 7 35

Figure 8 37

Figure 9 39

Abbreviations:

SMA = supplementary motor area proper

PSMA/pre-SMA= pre supplementary motor area

ACCr = rostral aspect of the anterior cingulate cortex

ACCd = dorsal aspect of the anterior cingulate cortex

LFP = local field potential

RP = readiness potential

IBE = inference to the best explanation

Introduction:

I. A primer on free will

It is a generally accepted and seemingly irrefutable fact that for many of our decisions, we act freely; we really *choose* when we decide whether to go to graduate school or work at a hedge fund, to go out with friends or stay in for the evening, to wear black or white socks for the day. We really choose, so it seems, these things in a way that we *don't* choose that our hearts beat, or that our fingernails grow, or that we yank back our hands after touching a hot stove. We really choose, so it seems, these in a way that trees that grow toward the sun don't, in a way that computers that make "choices" don't really choose. Call whatever this apparent capacity is, free will.

Now consider the following three statements:

(1) Humans have free will
(2) The universe is (for all intents and purposes)[1] deterministic
(3) Free will is incompatible with a universe that is (for all intents and purposes) deterministic

There seem to be initial reasons for believing each statement. (1) comes from introspection, over the sort of examples just described. If by free will, it is meant that decisions are to some extent *up to us*, that we have genuine control over which actions we take, then we certainly *feel* like we have it. (2) seems like a plausible empirical truth. If someone flips a coin twice and it lands heads one time and tails the other, few would deny that there is an explanation of this. Maybe he flipped it once with more power. Maybe a gust of wind came. Whatever the cause, there must be *some cause*. Coins don't, holding antecedents constant, magically land one way or the other. This seems to motivate determinism, which is the thesis that all events have causal antecedents that fully determine the future, such that some prior state of the universe in conjugation with the laws of nature entail everything that will happen. In other words, for any fixed past and fixed laws of nature, there is one

[1] The "all intents and purposes" caveat comes from recent developments in quantum mechanics that suggest our universe might actually be indeterministic, or be governed by probabilistic laws of nature. But if it's true that free will and determinism are incompatible with one another, then adding a few "coinflips" into the mix does not seem to provide us with any more freedom than what we have in a deterministic universe.

possible future; the universe, composed of just atoms bouncing around according to set laws, proceeds in a clockwork manner. (3) is an intuition held by many when thinking about what determinism would mean for human action. After all, if all events have causes, and our actions themselves are events, then our actions too have causes. Maybe an agent's decision to lift his right hand was caused by a conscious choice to do so, but presumably that conscious choice is itself an event with a cause. Even if the cause of that was another conscious choice, if determinism is really true, then we can keep tracking the line of causation back until we get outside of the agent (or before he was born). And this seems to strip genuine control from the agent; if everything is determined by past events being acted upon by the laws of nature, and nobody has control over the past or the laws of nature, then how does anyone have genuine control over their actions? If determinism is true, then it is true to say that billions of years before we were born, it was locked into the fabric of the universe that we would do each and every action that we have done, are doing, and will go on to do. If all of our actions were determined before we were born, then there is a certain sort of freedom that seems to be lost.

Although, given their initial plausibilities, we'd like to have (1), (2), and (3), there's a problem: we only get two. Suppose we pick (2) and (3). If the universe is functionally deterministic and this isn't compatible with free will, then we cannot have free will. This makes us hard determinists. On the other hand, we might want to hold on to free will (1). If we concede (3), we must reject (2), that our universe is functionally deterministic. This makes us libertarians (no relation to the political philosophy of the same name). Finally, if we pick (1) and (2), then we are conjoining the very two things which (3) says we cannot conjoin, so we must deny that free will and determinism are incompatible. That would make us compatibilists.

Call the debate on which number to slash as incorrect the free will debate. Here are two follow up questions. First, who cares? Second, can neuroscience do anything (or, perhaps, everything) to carry the debate forward?

II. Morality and legality:

To answer the first question, understanding the nature of human decision-making is far from just a complex empirical and philosophical curiosity; it seems to have deep implications in moral, legal, and clinical domains. In terms of *practical* outcomes, experiments suggest that a *deterministic* understanding of human decision-making (which rules out at least one understanding of free will) often makes people less likely to judge others as morally responsible (Nichols and Knobe, 2007), though findings have been mixed (Nahmias et al., 2005). Other findings suggest that people who disbelieve in free will are more likely to be aggressive and nonhelpful (Baumeister et al., 2009), and still others find that belief in determinism increases cheating behavior (Vohs and Schooler, 2008). Whether or not these are appropriate responses, better understanding human-decision making, then, has implications for both moral judgments and valuations as well as moral behavior.

More conceptually, in law, the notion of *mens rea*, or "guilty mind," as a requirement for successful conviction seems to place special emphasis on a person's intentions being relevant to their behavior. If our conscious intentions are discovered to be inefficacious in experimental tests, then this might begin to undermine our current understanding of responsibility in criminal courts (Greene and Cohen, 2004). Likewise, at the heart of the criminal justice system is the notion of retribution, the idea that criminals deserved to be punished for the simple fact that they made others suffer; that in doing wrong the criminal "owes a debt" to society and must suffer in order to balance out whatever scales were disrupted in a wrongful act (Loewry, 2009). Some believe that the idea of retribution relies on a conception of free will that is disprovable by science, and therefore better understanding the neural orchestration of our decisions could have enormous implications for criminal sentencing and criminal justice reform (Greene and Cohen, 2004). Finally, in the clinical

settings ranging from addiction to various mental health disorders, increased understanding of the mechanisms of self-regulation inherently hold potential for better diagnosis and therapeutic efficacy.

III. Cue Libet

The many developments in neuroscience over the past few decades have begun elucidating many of the processes of the mind in terms of physical processes of the brain, so it is seems all too natural to ask whether the phenomenon of free will can be dealt with in the same way. After all, arguing whether humans have free will seems to invite the empirical question of how our decisions are physically realized and under what circumstances. But while historically such answers have been largely intractable due to the physical and ethical barriers of directly studying the brain, recently neuroscience has begun to gain the technological capacity to begin investigating the physical implementation and realization of such behavior, presenting novel opportunities for mechanistic explorations of self-control. Through a multi-pronged approach involving lesioning studies (Laplane et al., 1977), electrical stimulation (Desmurget, 2009), and many other tools and approaches, neuroscientists have begun elucidating the functional areas and networks that orchestrate intentional movements. For example, researchers have identified the fronto-basal cortex as playing an important role in self-control decisions, and have discovered that the brain uses different pathways to realize intentional action compared to the inhibition of intentional action, the so called "veto" commands in volitional behavior (Brass and Haggard, 2007)(Schel et al., 2014).

The original studies directly related to free will stem from a paradigm developed by Benjamin Libet (Libet et al., 1983). Libet investigated the *temporal* relationship between conscious willing of actions, and the neurophysiological processes correlated with that experience. Participants repeatedly decided to freely flex their wrists while they watched a clock whose period was approximately two seconds. Participants were asked to remember where on the clock the dial was when they experienced freely choosing to flex their wrist, in order to get a time stamp of the

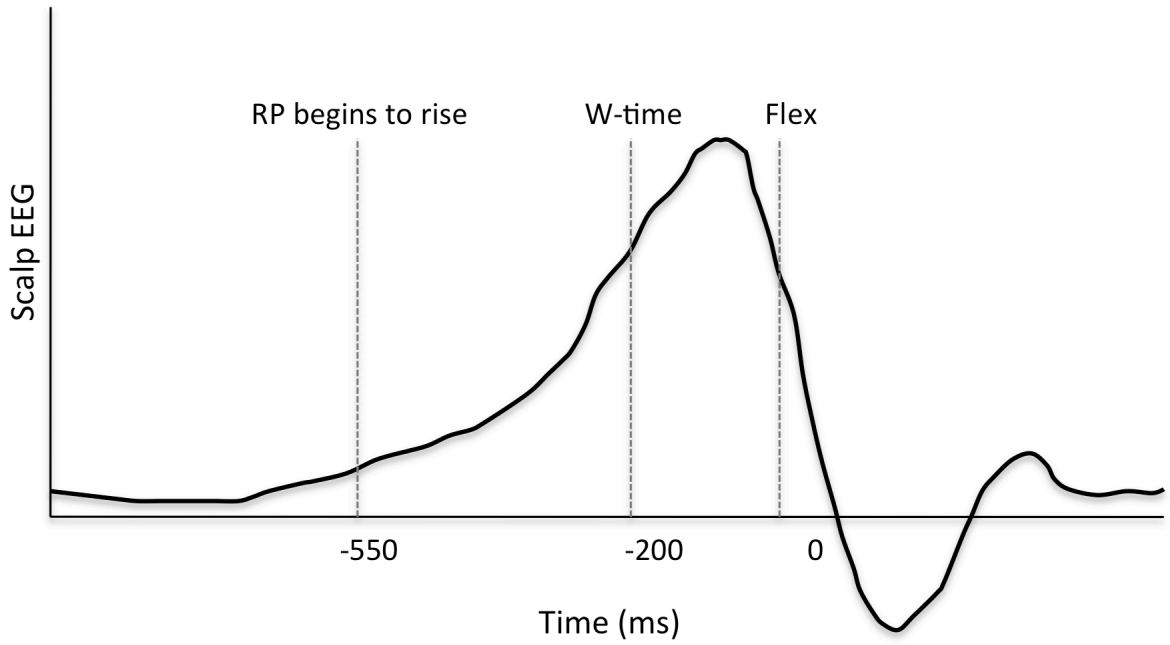

Figure 1 | The Bereitschaftspotential (Readiness Potential). Approximation of the original Libet finding, based on results from Libet (1983). A significant change in activity in the scalp EEG is detected in the time leading up to voluntary muscle movement, and this ramp up begins before participants' reports of consciously making the decision/being aware of their urge to move their wrist.

conscious decision. While this was being done, Libet also recorded electroencephalogram (EEG) activity from their skulls. After averaging activity from the blocks of 40 trials, Libet found that the conscious decision (dubbed W-time) to move reliably happened at about 200 ms prior to the wrist muscle movement. However, he also found that a reliable ramp up in electrical activity, the *Bereitschaftspotential* or readiness potential (RP), appeared reliably 550 ms before the muscle movement, and so 350 ms before the putative conscious decision (**Figure 1**) (Libet, 1983). Seeing the RP as the effective cause of whatever process leads to execution of wrist flexing, Libet and many have interpreted these results as showing that conscious wills have no role in causing actions in any robust way—since they appear after the (nonconscious) brain processes have begun, they come too late, and we cannot be free if our actions are the result of nonconscious, rather than conscious, processes.

The mere fact that conscious decisions, according to these studies, seem to temporally succeed brain activity has gained little traction among the philosophy community as discrediting various theories about free will (Mele, 2006)(Mele, 2009). This is for roughly two classes of reasons, one methodological and one conceptual. Methodologically, it is unclear whether the neural signatures actually proceed conscious decisions, or if instructions to be aware of conscious urges leads to systematic delayed reporting. Asking participants to be aware of their intention to move might introduce a lagging bias in reports of the conscious decision; there is no reason to believe that a participant accessing the content of the decision (being conscious of having decided) is simultaneous with consciously deciding. More specifically, it is an open question whether we need to be conscious of our conscious decision to act in order to consciously do it (Mele, 2009). Conceptually, it is unclear whether the neural correlates, even if temporally prior, are simply *necessary* conditions for decisions to be made, but not actually *sufficient* for specific decisions (and even if wrist flexes get ruled out, it is unclear whether the non-valenced, insignificant decisions related to wrist-

flexing are in any way analogous to the more valenced, significant decisions often associated with free will). Such criticisms of the original paradigm call for two specific improvements: first, if more robust data can be generated, that might quell any uncertainties about methodological deficits; second, if instead of correlations being found, *predictions* of decisions can be made, that might quell any uncertainties about whether the actual *content* of a decision is encoded in neural activity prior to the conscious will.

To try to provide both more significant and more predictive results, Soon and colleagues (2008) used functional magnetic resonance imagining (fMRI) to record from the frontopolar and parietal cortex regions while subjects were instructed to use one of two hands to press a button. The experimenters found that over 7 seconds prior to the conscious decision, which hand would be used could be predicted with better than chance accuracy (approximately 60%). While the weak predictive success might only reflect biasing of decisions, presumably some of the weakness is reflected in insufficient technologies and predictive algorithms—and a far greater predictive capacity, or so it is argued, would reflect deterministic neural processing rather than simple biasing.

IV. Intracortical recordings and the local field potential

Most of the standard methodologies of Libet experiments rely on extra-cranial recordings, such as EEG and fMRI, which have drawbacks compared to more invasive, *intracranial* recordings. For example, EEG has relatively poor spatial resolution, due to picking up large electrical activity generated by many areas of the brain. And fMRI, in sampling different slices of the brain and then interpolating and processing raw data to construct highly spatially resolute images, has relatively poor temporal resolution. While invasive recording concepts (such as spikes, local field potentials, and electrocorticography) remain far more rare due to ethical constraints, several labs have been able to co-opt the usage of electrodes implanted in patients for clinical practices, and conduct research (Engel et al., 2005)(Fried et al, 2011).

For example, following the classical Libet paradigm with single unit recordings in epileptiform patients with depth electrodes implanted for clinical purposes, Fried et al., (2011) measured spike activity patterns of neurons in various areas of the brain, including the supplementary motor area (SMA). They report neural recruitment at over -1,500 ms before the report of the conscious decision to move, in the form of both progressive increases and decreases in firing rates of the neurons in different brain regions, especially the SMA. Moreover, behavior of 256 SMA neurons could accurately predict the occurrence of the conscious decision with over 70% accuracy 700 ms prior to the conscious decision, as well as an approximate time point of the conscious decision (Fried et al., 2011).

However, far from acting as billions of independently spiking units, neurons often coordinate in networks, giving rise to much larger rhythmic potential changes than those of a given action potential. And while action potentials and neural spiking patterns have received the lion's share of attention during intracortical recordings, this belies the fact that voltage traces picked up from extracellular microelectrodes in fact consist of two main classes of superimposed signals: action potentials from single units and multi units (spikes) and slower potentials (**Figure 2**). While spikes occur at frequencies well above 300 Hz, they ride slower potentials in the 1-250 Hz range, local field potentials (LFPs), a massed signal. While the origin of the LFP is a matter of considerable debate, it is believed that the LFP reflects the electrical currents associated with this synaptic activity in local populations of units close enough to be detected by the electrode (Renshaw et al., 1940)(Mitzdorf, 1985). Since synchronous activity is sufficient for the generation of LFP, LFPs are therefore sensitive to *subthreshold* processes that are not discovered by examining action potentials; LFPs may represent summations of both excitatory and inhibitory dendritic activity, as well as other slower frequency modulations such as spike afterpotentials (Buzsaki, 2004).

Perhaps the greatest upshot of studying LFPs compared to action potentials is the capacity for greater signal decomposition. While action potentials are binary in nature and therefore confined to analyses across spike timing and rates, LFPs can be divided into different frequency bands, with such distinctions arising from characteristic oscillatory activity at band-limited components (**Table 1**). For example, theta activity has been implicated with cognitive activities like attention and thinking, while gamma rhythms seem like appealing candidates for modulating information transmission, more specifically with attention, such as spatial information processing during running (Ahmed and Mehta, 2012), forward transmission of stimulus information as well as suppression of other information to create selective attention (Adesnik and Scanziani, 2010), and even feature binding during conscious experience (Joliot et al., 1994). Generally, the higher frequency band signals (beta and gamma) have been associated with higher cortical processing, so are appealing targets for neural activity interrogation during Libet experiments. Thus, LFPs offer the possibility of adding another layer of richness to intracortical processing underlying volition. Moreover, since LFPs are a massed signal of groups of neurons rather than particular units, LFPs are more vulnerable to exploration of synchronicity, or coherence, in ways that would be far more noisy with spike-spike analyses.

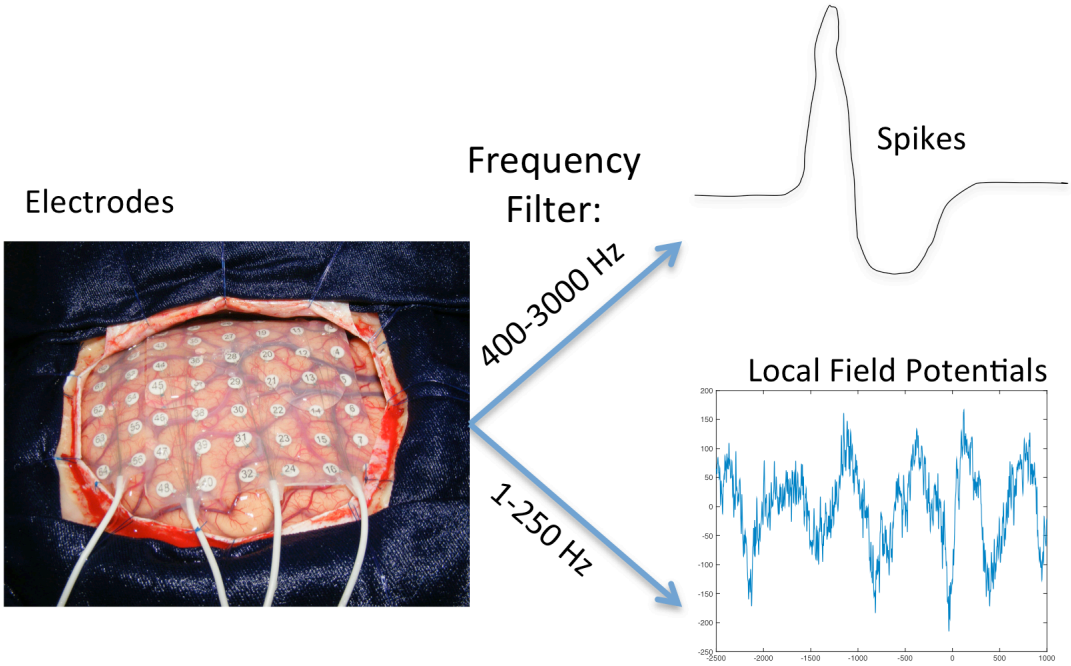

Figure 2 | Intracranial recordings constitute superimposed spikes and local field potentials. Spikes can be isolated by filtering at frequencies above 400 Hz, while local field potentials constitute the lower frequency recordings around the 1-250 Hz range. Image of depth electrodes from http://mnepilepsy.org/services/

Band	**Frequency Range (Hz)**
Delta	1-4
Theta	4-8
Alpha	8-12
Beta	12-25
Low Gamma	30-50
High Gamma	70-100

Table 1| Local field potential bands. LFP band frequencies are not clearly delineated; above table establishes thesis definition of frequency bands. Highlighted bands indicate bands associated with higher order cognition and volition, and the bands of interest for the course of this analysis.

Despite LFPs being one of the two main classes of intracortical signals, and despite the oscillatory characteristics of higher frequency bands being well associated with the cognitive processes likely to be underlying subject behavior during Libet tasks, to date there has never been a Libet-styled investigation of free will via LFP analysis. Because of the potential of the LFP to add another dimension to understanding the neural signatures of volition in human-decision making, the purpose of this thesis is to investigate the neural dynamics related to the LFP in the Libet task, to supplement the mechanistic understanding of human-decision making from non-LFP intra- and extra-cranial recording concepts. Specifically, we aim to 1) note baseline voltage trends that distinguish some baseline period with the time leading up to the conscious decision (pre-W time); 2) identify both broadband and band-specific LFP power (whether absolute or relative) signatures in the pre-W period; 3) identify region specific LFP trends; 4) explore interregional LFP relationships through any synchronicity they achieve through LFP-LFP coherence; and 5) use machine learning to see if the findings from 1-4 can be used to detect impending conscious decisions and in this way predict W-time before it occurs.

Methods:

Subjects:

Data presented in this paper are based on twenty-three recording sessions with eight patients diagnosed with pharmacologically intractable epilepsy. Since this thesis comprises a region-of-interest analysis on four predetermined regions, the data analyzed were pooled from the original study of twenty-eight recording sessions with twelve patients. Sessions and patients were excluded from analysis if no electrodes contained the regions of interest (one patient), or the task methodology differed from the following (three patients). Chronic depth electrodes were implanted for 7-10 days in order to localize seizure foci for possible subsequent surgical intervention.

We analyze the following four locations within the medial frontal lobe: the supplementary motor area proper (SMA), the pre-supplementary motor area (PSMA), the dorsal anterior cingulate cortex (ACCd) and the rostral anterior cingulate cortex (ACCr) (**3B**). While this thesis only spans the analysis of these regions, it should be kept in mind that the preliminary data included recordings from other regions as well, such as the temporal lobe. In fact, all electrode implantations were determined accordingly only to clinical criteria and for clinical purposes. As this thesis is a supplementary analysis on a previously generated dataset, further details about recording procedures, patient variability, and conformity to clinical guidelines and consent can be found as described previously (Fried et al., 2011).

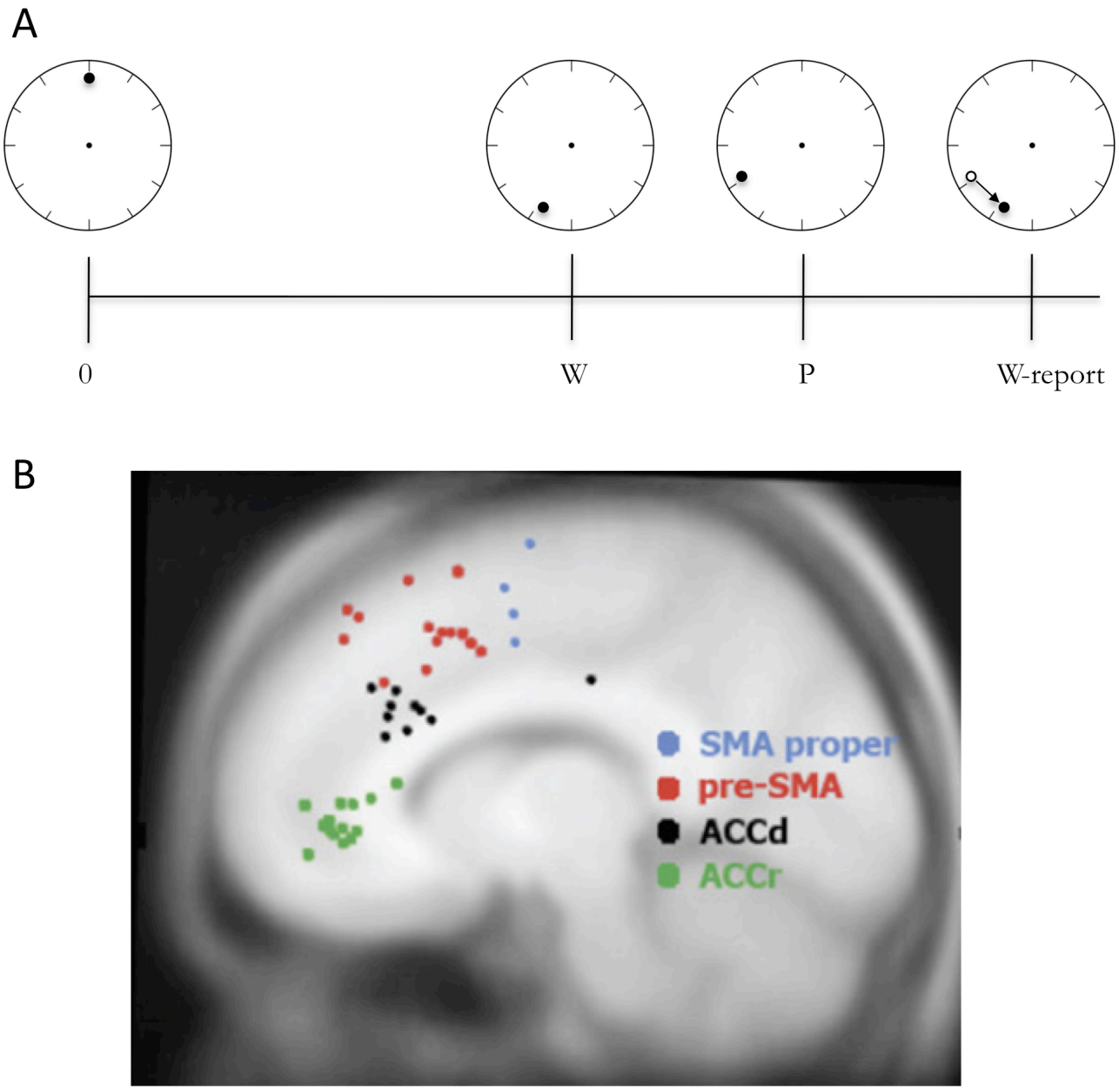

Figure 3 | Experimental paradigm and electrode placement
A. Schematic of Libet set up. At time=0, dial begins rotating around clock. After one complete rotation, participant waits for "urge to move" (W) and then presses the key (P). The dial then stops and the participant indicates where on the dial W-time occurred for a time stamp of the conscious decision to press the key.
B. Frontal lobe anatomical locations of electrodes in the four regions of interest. SMA, supplementary motor area proper; pre-SMA, pre-supplementary motor area; ACCr, rostral area of anterior cingulate cortex; ACCd, dorsal aspect of anterior cingulate cortex. Overlaid on a Montreal Neurological Institute brain (average of 305 brains) (Fried et al., 2007). Note: This figure was not generated by Garrett Lam. It was created by Itzhak Fried, Gabriel Kreiman, and Roy Mukamel.

Patients sitting upright in bed watched a computer screen featuring an analog clock with a period of 2,568 ms. For each trial, subjects were instructed to let the clock make one full rotation and then press the space bar of the laptop whenever they "felt the urge," bearing in mind where the dial was on the clock when they felt such an urge to press the button. Pressing the key (P time) caused the clock dial to stop turning, and subjects moved the clock dial back to where they remembered experiencing the decision to press the key. This moment of conscious decision is referred to as the time stamp of conscious free will (W time) (**3A**). Identifying W-time concluded the individual trial, and trials were repeated in blocks of 25. Trials were excluded if 1) W and P times were identical; 2) W time preceded P time by >1500 ms; 3) trial duration was >20 seconds; 4) subjects did not wait for a full rotation of the dial. Raw data provided in the dataset for this paper included 1,000 Hz sampling in the time periods of 2500 ms before and 1000 seconds after W time, as well as 2500 ms before and 1000 seconds after P time.

Local Field Potential Recordings:

Subjects were implanted with intracranial depth electrodes each containing nine microwires (eight recording, one reference). For data preprocessing, a notch filter was applied at 60 Hz, and the data were band passed from 1 Hz to 150 Hz in order to isolate the electrophysiological signal of interest, which throughout this text we refer to as the local field potential (LFP). In order to remove noisy trials and noisy channels and avoid potential artifacts, we treated as outliers and ignored from the following analysis any trials which fit either of the following criteria: 1) the amplitude of the LFP response (max(LFP)-min(LFP)) was greater than 3X the SD over all trials, or 2) the total power over the entire trial was greater than 3X the SD over all trials. Furthermore, the mean across all electrodes was subtracted from each LFP response at each time sample (common average reference) (Bansal et al., 2014).

Analysis of Neural Data:

All analyses were performed using MATLAB (Mathworks Inc., Natick, MA). All analyses were performed excluding the frequencies from [50-70] Hz in order to rule out possible artifacts from 60 Hz line noise from electrical outlets.

Power was compared between baseline and pre-W conditions, with baseline being the period from [-2500,-1988] with 0 marking reported W-time, and the pre-W time period being [-512,0]. A window of 512 ms allowed for the best combination of both frequency resolution and temporal resolution in subsequent analyses. For calculating the absolute power spectra ($\mu V^2/Hz$), a fast Fourier transform with multi-tapering was used, using the Chronux Matlab toolbox. This requires two parameters: the time-bandwidth product (TW) and the number of tapers (K), with the two providing spectral smoothing across frequencies. The number of tapers is often set to 2TW-1, and we used settings TW=3 and K=5 used for all relevant analyses (Ahmed and Mehta, 2012)(Jarvis and Mitra, 2001). For spectrograms, sliding windows of 512 ms were advanced with 100 ms window steps.

In order to assess the significance of the number of channels displaying LFP modulations in pre-W time in relation to baseline, we computed a distribution of the number of channels displaying significant LFP modulation for 1,000 iterations where we randomly shuffled the "baseline" and "pre-W" (or "non-baseline") tag labels in each trial. Shuffles of random label switching were also used in a similar fashion for both comparisons of power ratios at different frequency bands (**Figure 7**), and the cumulative distribution plots of coherence ratios at different frequency bands and across different regions (**Figure 8**).

For coherence analyses (Bansal et al., 2014), in each trial the coherence between two electrodes x and y at a given frequency f was calculated,

$$Cxy_f = \frac{\left|S_{xy}(f)\right|}{\sqrt{S_x(f)S_y(f)}}$$

where S_{xy} is the cross spectral density between the LFP time-series, which is normalized by the square root of the power spectral densities, S_x and S_y, of x and y, respectively. The coherence between two channels x and y in a given frequency band [f1-f2] Hz and period T be denoted

$$Cxy^{T}_{f1,f2} = \sum_{i=f1}^{i=f2} Cxy_i$$

In each of the electrode pairings, mean coherence across trials was defined as the following:

$$\bar{C}xy^{T}_{f1,f2} = \frac{1}{N_{trials}} \sum_{i=1}^{i=N_{trials}} {}_i Cxy^{T}_{f1,f2}$$

Then, to calculate the proportional coherence change (PC) in the pre-W time period with respect to the baseline period:

$$PC^{pre-W}_{f1,f2}xy = \left(\frac{\bar{C}xy^{pre-W}_{f1,f2} - \bar{C}xy^{baseline}_{f1,f2}}{\bar{C}xy^{baseline}_{f1,f2}} \right)$$

In order to train and test whether LFP activity could be used to discriminate and predict deviations from baseline activity as W-time approached, we used a support vector machine (SVM) (Hsu et al., 2003)(Hung et al., 2005) classifier to quantify LFP pattern changes before W. The machine learning algorithm maps training vectors into a higher dimensional space and finds a linear separating hyperplane that maximizes margin in this space (that is, best separates all the data from one class from all the data in the other class), and so generates a classifier that can classify at a single trial level. For discriminating between voltage traces different from baseline activity at some time period *t*, which demarks [t-512,t] ms prior to W, let the predictor values (i.e. absolute beta power)

for a single trial r in a single channel be defined ${}_rP_t$. Let the concatenated matrix P_t of predictor values for all the trials in that channel (r number of trials) be:

$$P_t = \left[{}_1P_t, {}_2P_t, \ldots {}_rP_t \right]$$

For a population of n channels, we assumed independent LFP frequency modulations and constructed a pseudopopulation (PP) vector by concatenating responses for each channel:

$$PP_t = \left[{}^1P_t, {}^2P_t, \ldots, {}^nP_t \right]$$

The baseline response of the predictor was defined as PP_{-1988}, i.e., the beta power from [-2500,-1988] ms prior to W. For training with multiple predictor types (z = number of predictor classes), we concatenated PP_t for all predictors to generate the classifying matrix (CM):

$$CM_t = \begin{bmatrix} PP_t^1 \\ PP_t^2 \\ \ldots \\ PP_t^z \end{bmatrix}$$

Thus, for any given time *t*, the input to the binary classifier with a label "+1" was CM_t and the baseline matrix (CM_{-1988}) was associated with the label "-1". The classifier was trained to discriminate between "+1" and "-1" examples, that is, whether based on the set of predictor types, baseline activity could be discriminated from the period of interest (at some particular time, on a single trial basis). We used a Gaussian/radical basis function (RBF) SVM kernel, which nonlinearly maps the samples into the higher dimensional space (Hsu et al., 2003). For training, we used as input a randomly chosen set of 70% of the trials and used the remaining 30% for evaluation of classification performance. Therefore, the data used to test the accuracy of the classifier were independent of and not seen by the classier during training and validation.

Results:

We analyzed a dataset of voltage trace recordings collected over twenty three recording sessions over eight epileptiform patients with pharmacologically intractable epilepsy. Since electrode placement was solely determined according to clinical criteria, electrodes were located in various areas over both frontal and temporal lobes; to increase statistical power using a region of interest analysis, we focused on only a subset of these electrodes from four regions in the medial frontal lobe: the supplementary motor area (SMA); the pre supplementary motor area (pre-SMA); the rostral part of the anterior cingulate cortex (ACCr); and the dorsal part of the anterior cingulate cortex (ACCd). We focus on these regions over the others as they have been implicated in volition and previously associated with spike—rather than LFP—modulation in the identical Libet task (Fried et al., 2011). Patients participated in an analogue experiment from the original one developed by Libet (1983): subjects watched the center of an analog clock with a period of 2568 ms on a laptop, and were instructed to freely press a laptop key after one rotation of the dial, bearing in mind when they first became conscious of the urge to move. After pressing the key (P-time), they indicated when they had this conscious experience (W-time), which provided a time stamp for the conscious will.

We report an analysis of extracellular activity recorded from a total of 426 channels in the medial frontal lobe: 130 from the SMA; 126 from the pre-SMA; 104 from the ACCr; and 66 from the ACCd. **Figure 2** shows an example single trial voltage trace (bandpassed for LFP frequencies).

After preprocessing and filtering the data (**Methods**), in order to determine whether channels altered in their LFP activity in relation to W-time, the reported conscious decision to press the key, we aligned the voltage traces in relation to W. For the power spectra, in order examine differences between "baseline" activity and LFP activity just prior to W, we decomposed the signal into two time windows of 512 ms in duration: a baseline window (-2,500 to -1988 ms with respect to W); and a pre-W window (-512 to 0 ms with respect to W). The baseline was selected as the earliest

possible reference of activity, in line with previously used baseline start points in Libet experiments (Fried et al., 2011). A 512 ms window duration provided the best combination of temporal resolution (to calculate frequency band modulations over time) and frequency resolution (to discriminate the power at close frequency values). Using a multi-tapering fast Fourier transform with a time-bandwidth product TW=3 and tapers K=5 (in order to provide spectral smoothing, see **Methods**), we calculated the total absolute LFP power at each frequency band as the sum of individual frequency powers within the defined band (**Table 1**). For spectrograms, the same multitapering was used with a window step of 100 ms. **Figure 4** shows example power spectra and spectrograms of individual channels, averaged over all trials. The channel shown in **4A** and its associated spectrogram in **4B** shows a striking increase in high gamma power activity in pre-W time compared to the baseline ($p=1.9x10^{-6}$, Wilcoxon rank-sum test), while a different channel's power spectrum (**4C**), showed a decrease in beta power (p=.006), particularly as W-time approached (its associated spectrogram in **4D**). These channels undergo LFP modulations substantially different from baseline activity well after the baseline period, but before the subjects became aware of the decision/urge to move.

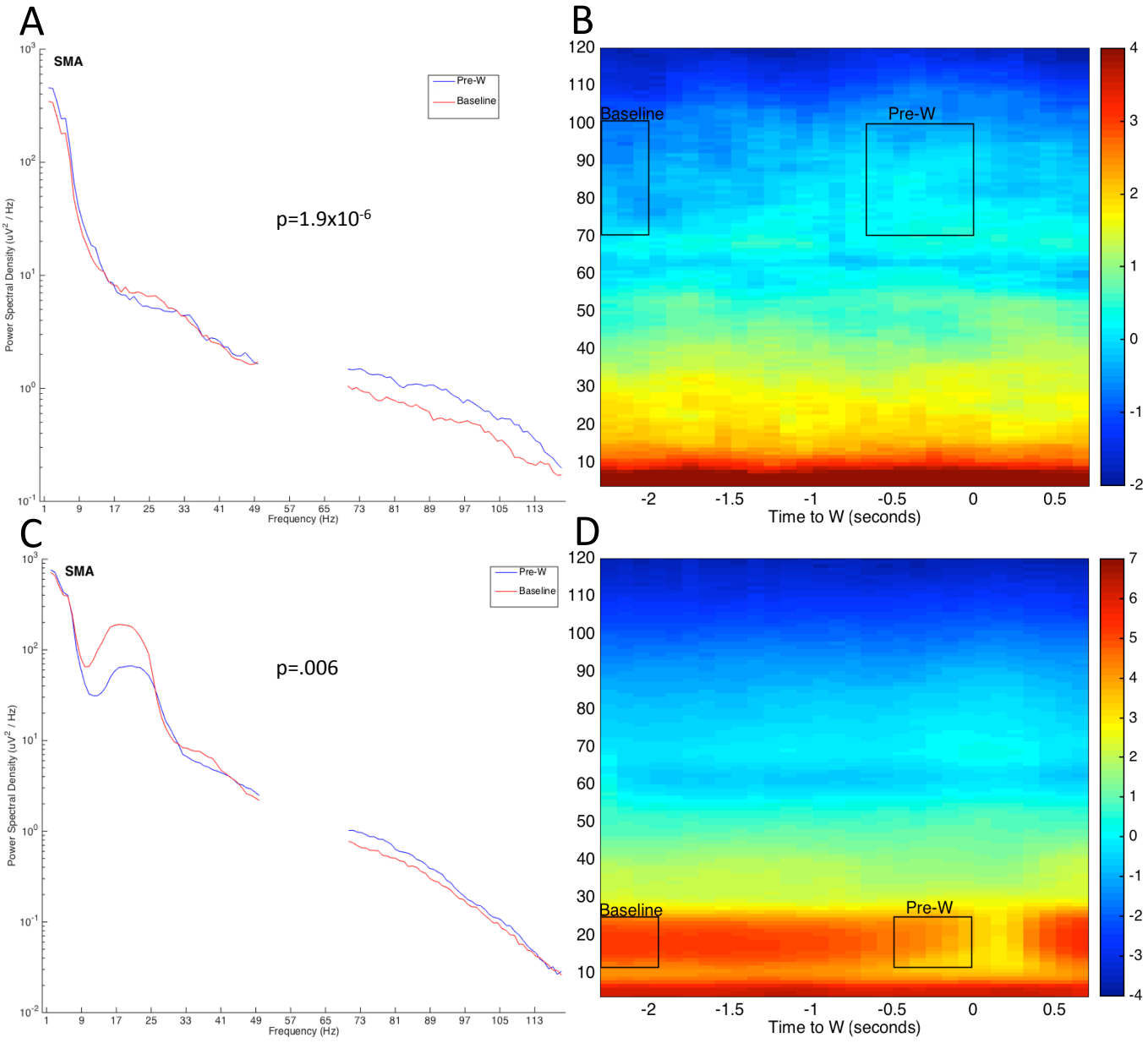

Figure 4 | Example power spectra and spectrograms of significant channels. The band passed and filtered waveform of each channel was decomposed into two temporal windows: 1) baseline (-2,500 to -1988 ms with respect to W), and pre-W (-512 to 0 ms with respect to W). Spectral power was calculated using a fast Fourier transform with a time-bandwidth product TW=3 and tapers K=5 for determining smoothing across frequencies (**Methods**). Example channels showing significant differences in the high gamma (70-100 Hz) frequency band (Subject 1, left SMA) (**A**, $p=1.9x10^{-6}$, Wilcoxon rank-sum test) and in (**C)** the beta (12-25 Hz) band (Subject 4, left SMA) (**C**, P=.006), with associated spectrograms in **(B)** and **(D)**, respectively. The channel in **A** and **B** detected a tonic increase in high gamma power in the pre-W period compared to baseline; the channel in **C** and **D** detected decreases in beta power in the pre-W period. For time period power spectra, activity from 50-70 Hz is ignored (so that 60 Hz noise is not considered). Note that spectrograms were generated with a sliding window of 512 ms and a window step of 100 ms, so the first 256 ms of baseline activity in the spectrogram is truncated.

In order to assess the region specific and frequency specific characteristics of any significant LFP modulations in pre-W, we repeated the same analysis as in **Figure 4**, pooling significant channels by region. Due to their previous associations with cognitive processing and volition, we decided to focus on three of the higher frequency bands: beta (12-25 Hz), low gamma (30-50 Hz), and high gamma (70-100 Hz)—though broadband activity from 4-150 Hz, excluding contributions from 50-70 Hz (see **Methods**), was also calculated for reference. Comparing LFP activity in the pre-W time period compared to baseline, we found that 46 out of 426 channels in the medial frontal lobe (10.8%) significant changed in absolute high gamma power (Wilcoxon rank-sum test, $p<.01$), well above levels expected by chance (**Table 2**).

The most pronounced areas of high gamma modulation were the SMA and pre-SMA, where respectively 18 out of 130 (13.8%) and 17 out of 126 (13.5%) of channels displayed significant changes. Those in the rostral and dorsal aspects of the ACC (3.2% and 10.6%), were also higher than chance. Significant numbers of channels were also observed in the beta frequency band in both the pre- and proper SMA, as well as the rostral ACC. By contrast, there was far less modulation in the low gamma frequency band, with no significant channels in either the SMA or pre-SMA (**Table 2**). Decomposing the LFP signal into these frequency bands appeared to aid in uncovering LFP modulations; in every region, the number of channels displaying broadband significant activity was less than or equal to the number of frequency band specific channels in at least one of the concentrated bands. Sampling absolute LFP power across such a wide spectrum likely implicates noise from unmodulating LFP frequencies, obscuring biologically significant signals as those found in the beta and high gamma bands. Because of such activity being confined to beta and high gamma, we confined the remainder of our power analyses to those two frequency bands (**Table 2**).

	SMA Proper (n=130)	PSMA (n=126)	ACCr (n=104)	ACCd (n=66)
	Significant Channels, Absolute Power			
Broadband (4-150 Hz)	10 (7.7%)	6 (4.0%)	5 (1.9%)	0 (0.0%)
Beta (12-25 Hz)	15 (10.8%)	6 (4.8%)	5 (3.8%)	0 (0.0%)
Low Gamma (30-50 Hz)	0 (0%)	0 (0%)	3 (1.0%)	3 (4.5%)
High Gamma (70-100 Hz)	18 (13.8%)	17 (13.5%)	4 (3.2%)	7 (10.6%)

Table 2 | Anatomical distribution of responses in regions of interest (pooled from 8 subjects). Total number of channels in region, frequency band of test, number (and percentages) of channels in each region displaying significant differences (Wilcoxon rank-sum test, $p < .01$) in absolute power spectra between baseline and pre-W time at specified band. SMA, supplementary motor area proper; PSMA, pre-supplementary motor area; ACCr, rostral area of anterior cingulate cortex; ACCd, dorsal aspect of anterior cingulate cortex.

While the absolute power of the LFP measures the sum of frequency specific power values over a given frequency domain, the metric can fail to discriminate changes in the relative contributions of each frequency band to the total signal in any given time period. For example, power in the high gamma band might stay unchanged in the pre-W and baseline time periods, but a tonic decrease in absolute broadband LFP power as W approaches would implicate high gamma as increasing its proportional contribution to the total power signal at the cost of decreasing power in the frequency bands. Since relative power spectra have also been used to compare band specific power across time (Cardin et al., 2009), we analyzed the relationship of relative LFP power as well, though the analysis did not result in any significant channel modulations over and above those seen in the anatomical distribution of responses seen in **Table 2**.

In order to determine whether such proportions of channels displaying significant (Wilcoxon rank-sum, $p<.01$) deviations from baseline activity in the absolute LFP were robust, we computed a distribution of the number of channels displaying the same degree of LFP modulation where we randomly shuffled the "baseline" and "pre-W" tag labels in each trial. We repeated this shuffle for 1,000 iterations and compared the distribution of shuffle values to the actual experimental values for beta and high gamma frequencies. In both frequency bands in each region except for beta activity in the dorsal aspect of the ACC (**5G**), the experimental number of significant channels was significant ($p\leq.001$ in all remaining regions and frequency bands) (**Figure 5**).

While absolute LFP power in both beta and high gamma frequencies, then, is significantly different in the pre-W period compared to baseline, this does not color in the temporal window between the defined baseline and pre-W periods—how does the number of significant channels detecting such activity *change* as W is approached? Since the timestamp for the conscious decision is subjective in nature, it is likely fallible to inaccuracy, and thus has been a source of considerable criticism toward the robustness of Libet studies. Given that 1) there is no guarantee that

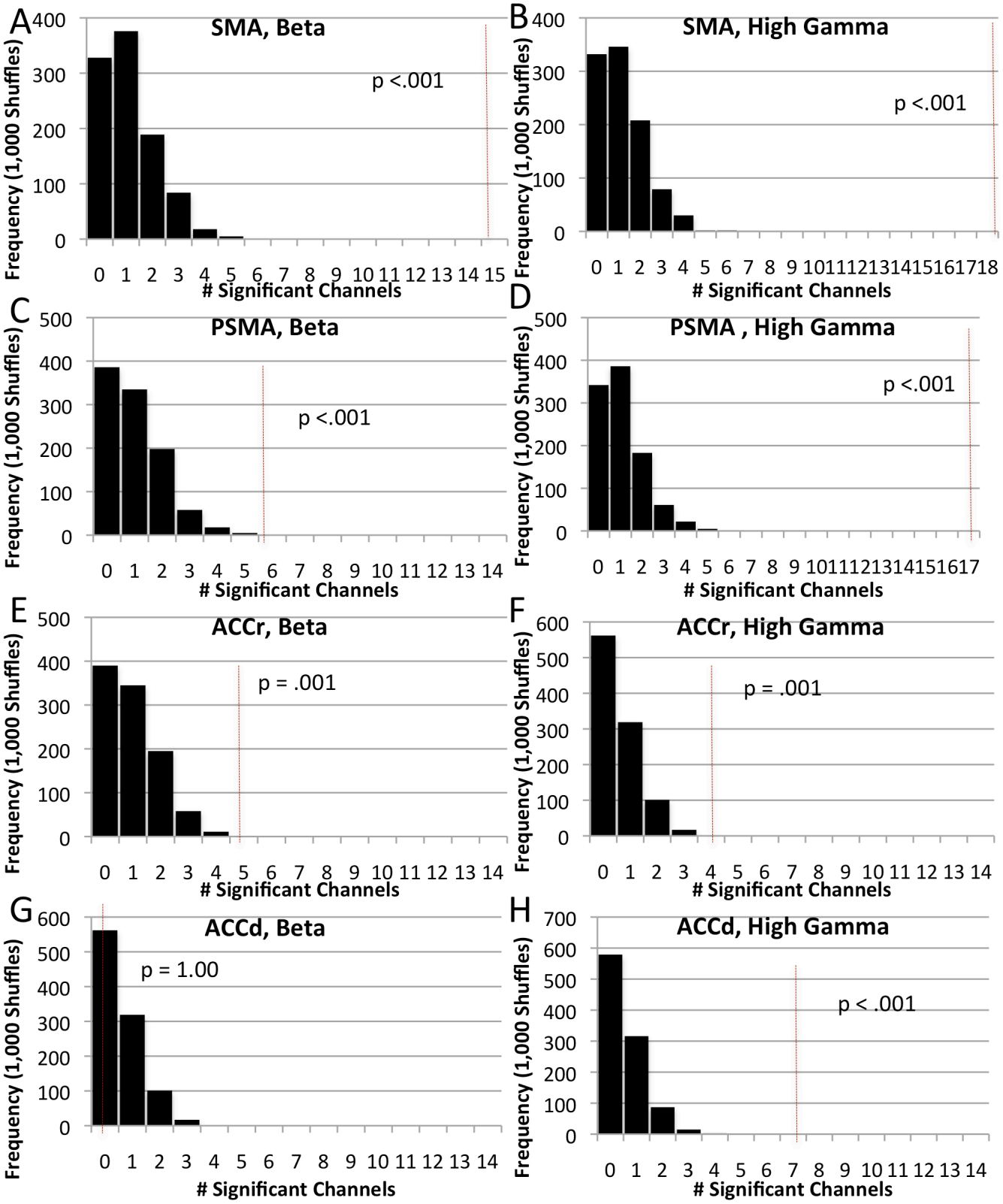

Figure 5 | Number of modulated channels are significant compared to shuffle. To assess the significance of the number of pre-W modulating channels described in **Table 1**, we computed the distribution of significant channel values in 1,000 iterations where we randomly shuffled "baseline" and "pre-W" object labels in each trial. Comparisons between the shuffled distribution and the actual number for the region (red dotted line) in beta and high gamma bands (**A** and **B**, respectively) in the SMA; in the PSMA (**C** and **D**); in the ACCr (**E** and **F**); and in the ACCd (**G** and **H**).

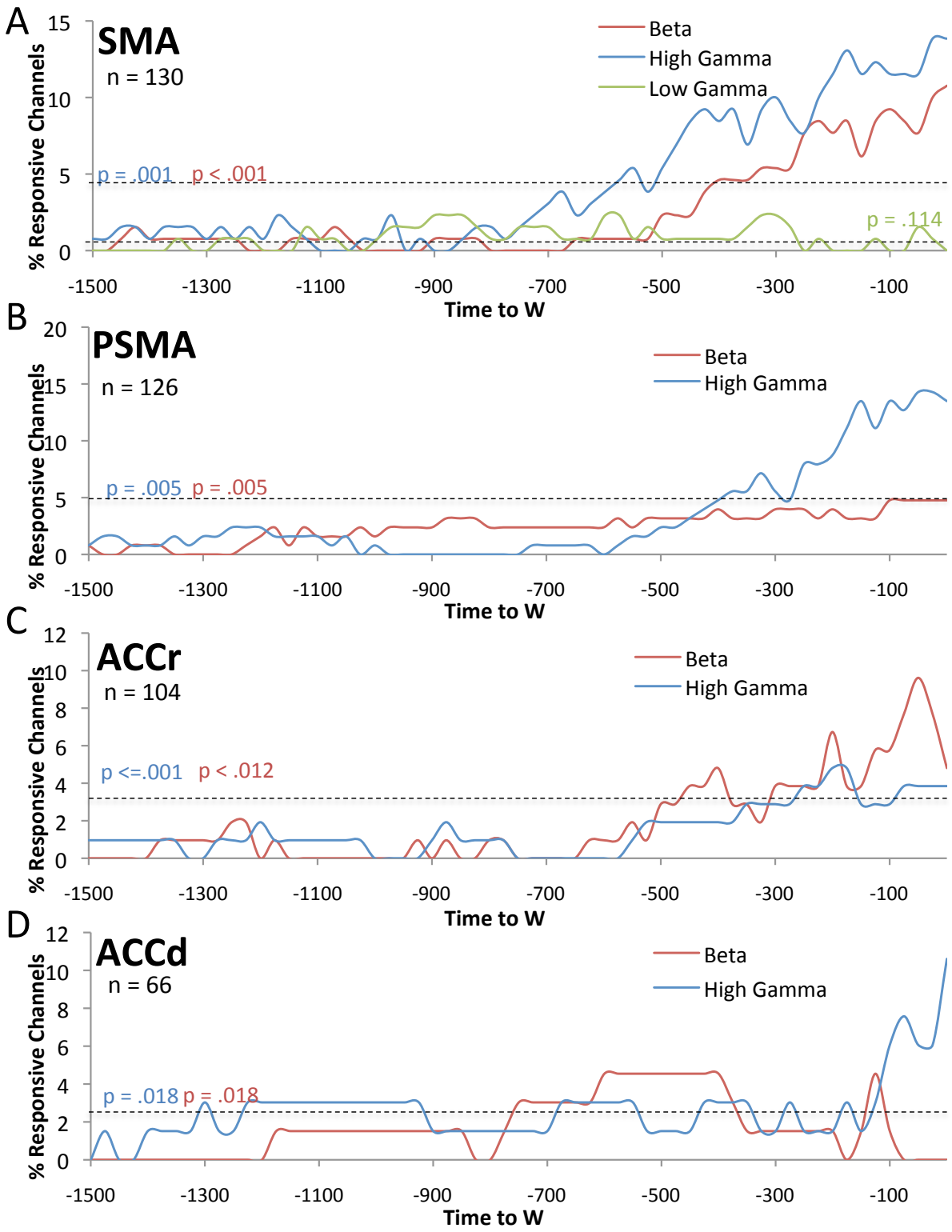

Figure 6 | Progressive recruitment of modulated channels as W is approached. Percentage of channels displaying significant changes in absolute LFP at beta and gamma frequency bands (Wilcoxon rank-sum) as a function of time before W, in SMA (**A**), PSMA (**B**), ACCr (**C**), and ACCd (**D**). For each channel, baseline LFP power was [-2500,-1988] ms relative to W; we then calculated the same in a 512 ms long sliding window (25 ms steps) from times -1500 ms to 0 ms (as the end point in the window) and assessed significant channels from baseline. The dashed black line indicates threshold for significant activity based on histogram shuffles in **Figure 5**. Insignificant activity is shown in the low gamma band for **A**, and excluded for clarity in remaining regions.

participants' awareness of their urge to press the button is accurate, 2) the cognitive process of becoming aware of consciously deciding is quite possibly distinct from (and so temporally consequent to) actually consciously deciding, reports of W-time may be systematically biased and lagged. If this is the case, then "significant" modulations in the pre-W time period might actually reflect brain activity concurrent with or even consequent to the real W-time, stripping Libet experiments of their original thrust: having neural activity antecedent to conscious decision making.

Since calculating period differences from baseline in the temporal windows between baseline and the defined pre-W is conceptually identical to temporally shifting W-time backwards, we redefined the pre-W time period by a fixed amount ranging from 0 to -1500 and repeated the previous analyses to compute the number of significant regional channels at a given frequency band, using steps of 25 ms (**Figure 6**). We found that across all regions (**6A-D**), there was a gradual recruitment of significant channels in the high gamma frequency band, and this pattern was also observed for beta activity in the SMA (**6A**), the PSMA (**6B**), and the ACCr (**6C**). In fact, constructing temporal profiles of channel recruitment also revealed significant deviation in beta activity prior to the original pre-W time period from the window of approximately -1000 to -500 ms prior to reported W-time in the ACCd (**6D**). Since this modulation extinguished before the standard pre-W period, it was lost in the analysis only considering the original pre-W time, but suggests antecedent significant activity in the ACCd even 1,000 ms prior to W-time. Since low gamma (30-50 Hz) modulations also might have occurred prior to the standard pre-W time period in the same was as ACCd beta modulation, we repeated the analyses for low gamma, but found no significant activity consistent with any temporal shifts (**6A** for insignificance overlaid, not shown for other regions).

Thus, significant differences between time periods prior to W and the baseline are consistent with shifts of W-time of even greater than 250 ms in some regions, mitigating worries of lagging or systemic delay W-time reports so long as the degree to which participants systematically err in their

reports of W-time does not exceed the degree to which shifts in W-time to compensate that lag (i.e. 250 ms in some cases) does not compromise the significance of LFP modulation.

While the previous analyses affirm significant deviations in beta and high gamma activity as W-time is approached, they add little color to the nature of these changes, such as whether there is a general shift toward increasing or decreasing power in the pre-W period, or whether channels in a given region move bidirectionally. To distinguish between these possibilities, for each channel in a given region we calculated the *ratio* of absolute band power in the pre-W time period over that in the baseline (the log of the final ratios was taken to remove bias). Comparing the distribution of actual power ratios to random shuffles where the power labels were randomly assigned, we found that in the pre-W time period, compared to baseline, there is generally a decrease in absolute beta power and an increase in absolute high gamma power (**Figure 7**). This trend was reproduced in both the SMA (**7A** and **7B**) as well as in the pre-SMA(**7C** and **7D**), with beta also being left skewed in the ACCr (**7E**) and high gamma also being rightly skewed in the ACCd (**7H**). By contrast, the rostral aspect of the ACC was the only region in which high gamma activity was skewed leftward (**7F**).

Thus, various areas within medial frontal lobe regions do not appear to bidirectionally change in absolute LFP power with respect to W-time, as single and multi- unit activity has been found to increase or decrease in firing rate in the same regions during Libet tasks (Fried et al., 2011). By contrast, there seems to be a tonic diminishment in beta power while there is an elevation in high gamma power.

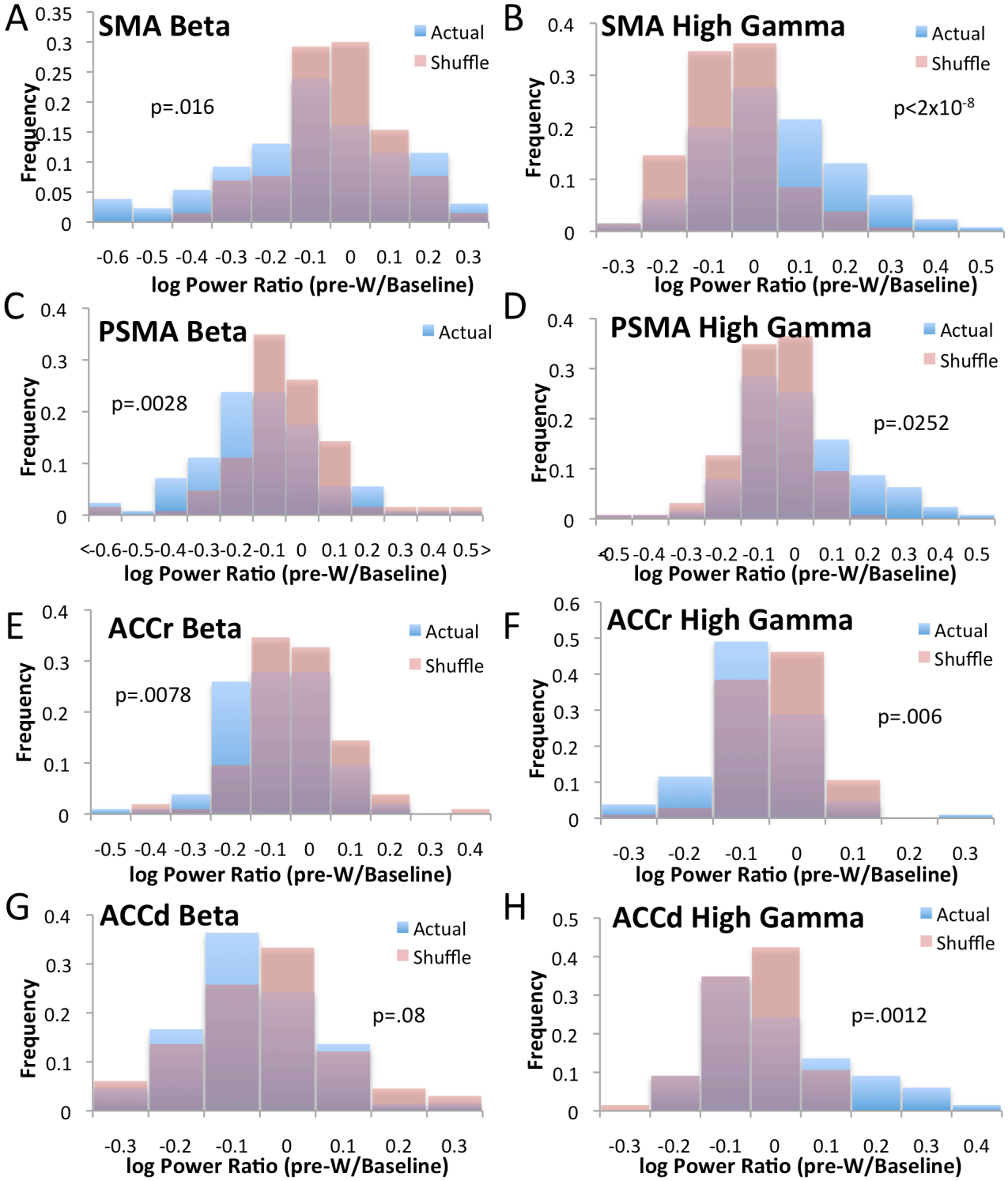

Figure 7 | Directional beta and gamma power correlations to deviations from baseline. To assess whether significant changes in LFP power were directional in nature, the distrubtion of power ratios between the pre-W and baseline periods were calculated for each region (blue, actual distributions; red, shuffled distributions by randomly shuffling object labels (**Methods**)). (**A**,**B**) Distribution of power in the SMA is skewed leftward in the beta band (p=.016, two-sided t-test of unequal variance) and rightward in the high gamma band ($p < 2x10^{-8}$). Similar results were found in the beta band for **C** the PSMA (p=.0028), and **E** the ACCr (p=.0078). Similar results were found in the gamma band in **D** the PSMA (p=.0252) and **H** the ACCd (.0012). By contrast, high gamma power appears to decrease as W time approaches in **E** ACCr (.006).

Up to this point, LFP analysis has only seemed to extend the same type of story portrayed by examining spiking patterns in the Libet tests: neural antecedents are found in the pre-W time that significantly differ from baseline. However, a dominant advantage of local field potential analysis over spiking in terms of elucidating the when, where, and how of neural communication lies in its greater vulnerability for interregional interrogation. Since LFPs are a massed signal reflecting multiple synaptic dendritic inputs over a much larger area than the spatially and temporally specific firing patterns of a single unit, coherence calculations—which reflect the linear coupling or degree of "synchrony" between two signals—between spikes and LFP, or between two LFPs, are far more robust than spike-spike interactions. Moreover, coherence can be normalized by responses in each electrode and so can both be unaccounted for by enhanced or diminished power *and* manifest without enhanced or diminished power in either electrode—it therefore offers a layer of understanding neural interactions unavailable to sole spike pattern analysis.

To evaluate any potential such interactions, we computed the degree of coherence between the signals derived from electrode pairs (**Figure 8**). To focus on interregional connections and to avoid potential synchronization due to common average reference, we restricted our analyses to all pairwise channel interactions where two channels from some permuted couple of the four regions of interest came from different regions (**Methods**). The most prominent interregional relationship existed between the SMA and the pre-SMA, where there was a significant increase in coherence in the pre-W period compared to baseline in the high gamma frequency band (**8B**, $p=.007$, two tailed t-test of unequal variance), concurrent with a diminishment in beta coherence (**8A**, $p<2x10^{-9}$). While no other regional permutations manifested coherence modulations at both frequency bands, the dorsal and rostral aspects of the ACC increase in their beta synchronization in the pre-W period (**8C**, $p<7x10^{-4}$). Moreover, we even found one relationship between the motor area and the anterior

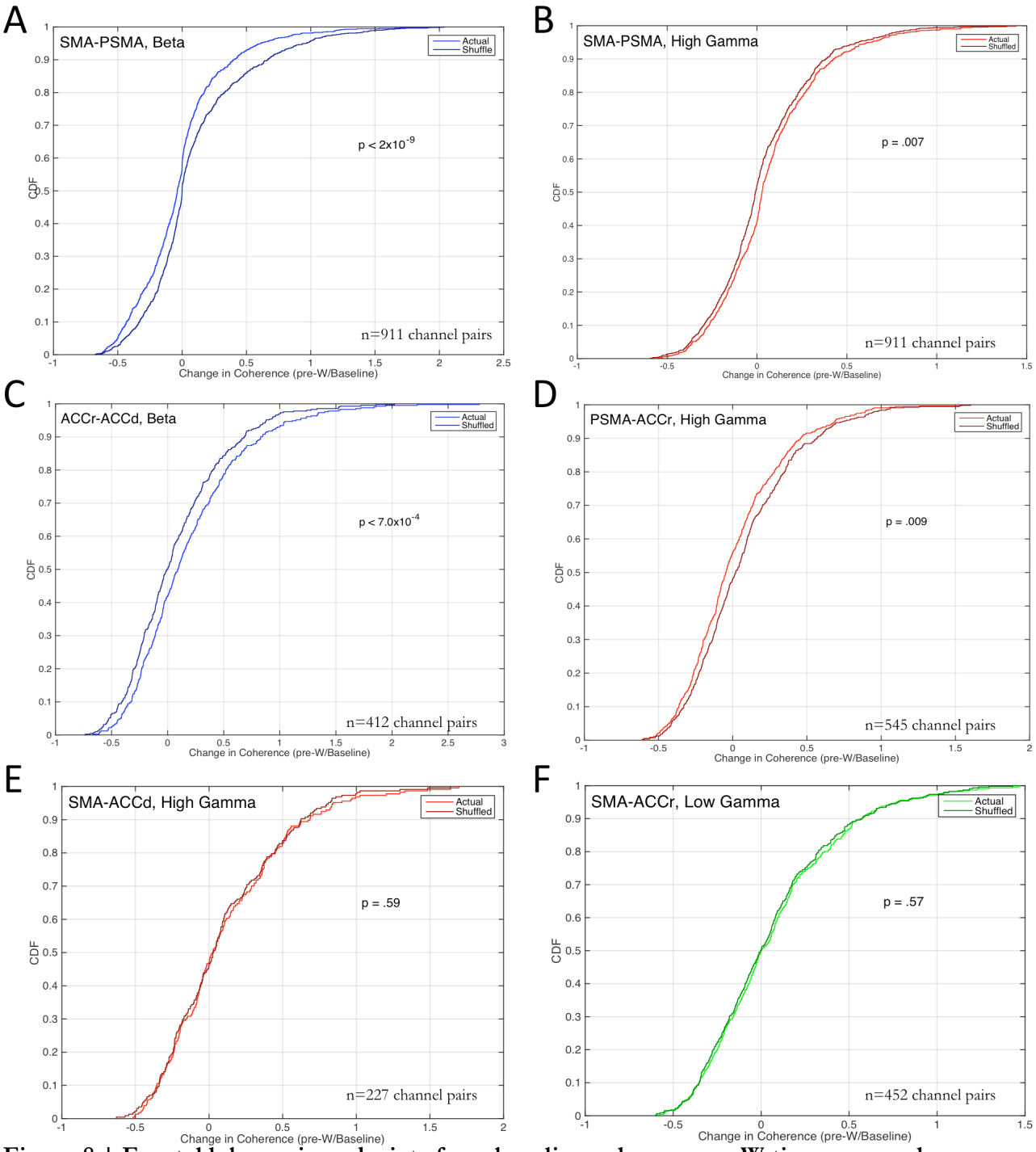

Figure 8 | Frontal lobe regions deviate from baseline coherence as W-time approaches.
To supplement intraregional explorations of significance, interregional relationships were examined by calculating coherence across regions and comparing to a shuffle (**Methods**).

A. W-time marks diminished coherence between SMA and pre-PSMA areas in the beta frequency band (two tailed t-test of unequal variance, $p < 2x10^{-9}$).

B. High gamma coherence is elevated in pre-W time between SMA and pre-SMA ($p = .007$).

C. Rostral and dorsal ACC manifest elevated coherence in pre-W time ($p < 7x10^{-7}$).

D. PSMA and ACCr manifest diminished high gamma coherence in pre-W period ($p = .009$).

E,F. Example regional interactions without significant changes in distribution of changes in coherence between pre-W and baseline periods. Note tighter grouping of actual and shuffled distributions. All other permuted interregional coherences across beta, low gamma, and high gamma were insignificant.

cingulate cortex, in the form of desyncrhonization of high gamma between the pre-SMA and the rostral aspect of the ACCr (**8D**, p.009). As with the temporal profiling of absolute LFP power, gamma synchronization did not change significantly across all trials in any of the regional permutations (**8F** for example insignificant gamma coherence changes).

Given these various neural signatures correlated with departure from baseline and the impending W-time, we hypothesized that the conscious decision to press the laptop key would be dependent on some ensemble of neurons whose activity could be used to discriminate W-time and in this sense "predict" the impending conscious decision; thus, we asked whether we could decode the departure from baseline on a trial by trial basis. This seems to be particularly important for relevance to free will, as the original *bereitschaftspotential* and other Libetian-styled findings rely on *averages* of activity over multiple trials, while volition obviously takes place at a "single trial" level.

To try to decode W-time at a single trial level, we used a support vector machine (SVM) classifier (Hung et al., 2005). Focusing on the SMA, we constructed a pseudopopulation considering channels from the SMA proper across all sessions and subjects. For each predictor type (i.e. beta power), we concatenated the value for each trial in each channel. The baseline response was the concatenated matrix of all predictor classes (i.e. absolute beta power, absolute gamma power, etc…) in the baseline time period, while the pre-W response was the same collection of predictor types for the time period prior to W (or, for the temporal profile of classification seen in **Figure 9**, any non-baseline time period). Thus, at any given time, the input to the binary classifier with a label "+1" was the matrix with the non-baseline predictors, while the class outcomes associated with the label "-1" were those from the baseline. The classifier was trained to discriminate between "+1" and "-1" examples, that is, whether based on the set of predictor types, baseline activity could be discriminated from the period of interest (at some particular time, on a single trial basis). We used a Gaussian/radical basis function (RBF) SVM kernel. For training, we used as input

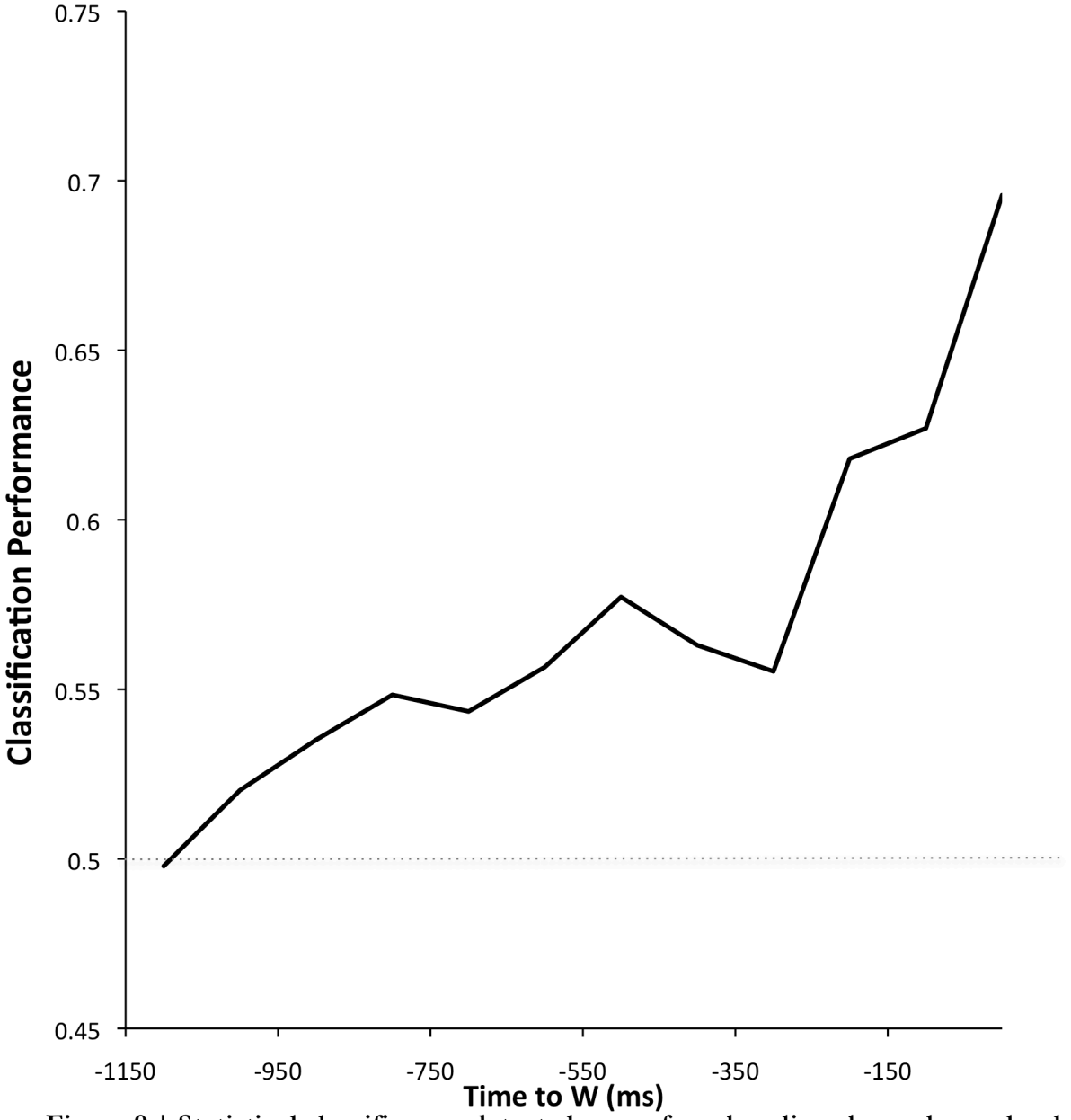

Figure 9 | Statistical classifier can detect changes from baseline above chance levels. Performance of the support vector machine (SVM) in discriminating between LFP signatures in the baseline period compared to some period before W. We constructed a pseuodopopulation of the 18 channels in the SMA manifesting significant high gamma absolute power deviations from baseline, and assigned the response of four predictor classes: absolute high gamma, beta, and gamma power, as well as relative high gamma power, as belonging in either the baseline [-2500, -1988] or non-baseline [t-512,t] period, where t=Time to W (x axis). The y axis shows the performance of the classifier against the dotted black line—expected performance by chance. A Gaussian/radial basis function (RBF) classifier was used; performance is calculated by the mean performance of the SVM at a given time window across a rudimentary grid search parameter optimization with 10 cross validation. 70% of trials were used for training and validation, while the remaining 30% were used to test the classifier; thus performance was tested with independent data not used as input into the classifier for training. A pseudopopulation of only 18 channels can discriminate between baseline and pre-W activity with accuracy greater than chance.

a randomly chosen set of 70% of the trials and used the remaining 30% for evaluation of classification performance. Therefore, the data used to test performance of the classifier was not seen by the classier previously during training. We found that the ensemble of 18 channels displaying significant absolute high gamma modulation in the pre-W time period compared to baseline in the SMA was sufficient for classification higher than chance several hundred ms prior to W-time (**Figure 9**), based on a concatenated input matrix of four predictor types: absolute high gamma power; absolute gamma power absolute beta power; and relative high gamma power.

Discussion: Empirical Reflections

We have presented evidence that before the conscious decision to press a key (as reported by participants), small ensembles of neurons, as detected by their massed contributions to local field potential signals, exhibit significant modulations in their activity that not only precede conscious volition but can also predict volition at levels well above chance, and make such predictions on a trial by trial basis.

Regarding the specific regions implicated in volition, our findings of the greatest number of significant modulated channels being in the SMA proper and pre-SMA is consistent with prior findings about their relative importance in volition (Fried et al., 2011). And regarding the previously unexplored domain of local field potential activity in those areas during Libetian studies of volition, we report that the beta (12-25 Hz) and high gamma (70-100 Hz) LFP frequency bands appear to be those most important in regulating volition, particularly as they operate in the pre-SMA and SMA proper. Indeed, in both of those areas, the time period prior to W exhibits a significant decrease in beta power in parallel with an increase in high gamma power, as compared to the baseline period.

Our investigations underscore the advantages of depth electrode investigations of cognitive tasks. In many other non-invasive forms of neural activity monitoring, a trade-off exists between spatial specificity and temporal resolution. For example, EEG recordings have robust temporal resolution but lack in spatial resolution, while the BOLD signal is difficult to decompose into pinpoint time points of activity. Taking advantage of intracranial electrodes used for clinical purposes allowed for precise brain region analysis without compromising temporal resolution. This is particularly salient for Libet studies due to the entirely subjective nature of the key variable: W-time, which by necessity is reported and quite likely fallible to extensive error. The temporal resolution of depth electrodes allowed us to rerun analyses under the assumption that W-time was misreported (by shifting it earlier), and our findings—that the degree of significant modulation in

beta and high gamma frequency bands is consistent with W-time being systemically lagged even hundreds of ms—deflate, if not diffuse, the worry that such apparently antecedent modulations are actually not occurring prior to the conscious decision.

Our study also underscores the advantages of local field potential analyses. While the lion's share of intracranial analyses often focus on spike timing patterns, lower frequency bandpassing into the LFP domain provides a significant upshot, particularly in terms of coherence studies. If spike-spike coherences reflect the unlikely event that a given Harvard student knows a given Yale student, then LFP-LFP coherence reflects the more likely possibility that someone from a group at Harvard knows someone from a group at Yale. Thus, we were able to report the first ever significant changes in coherence in Libet studies. Namely, between the SMA proper and the pre-SMA, there is a dissipation of synchrony at the beta band as W-time approaches, concurrent with an increase in high gamma coherence. Recall that beta power in the pre-W time period compared to baseline also decreased in *both* the SMA proper and the pre-SMA, and that high gamma power increased in *both* regions as W-time approaches. It seems plausible, then, that the increase in high gamma coherence between the two regions triggers some neural feedback leading to an increase in absolute high gamma power in both; and likewise, a decrease in coherence among beta waves between the two regions leads to a tonic decrease in beta power in both.

Perhaps the greatest downside to such depth electrode activity was the associated small sampling of activity, both in terms of number of participants used and areas of the brain analyzed. Since the availability of subjects was contingent on clinical justifications and since placement of electrodes was solely determined by clinical criteria, our sample of actual trials was smaller than would be ideal, and our sample of recording locations was far from complete. Many areas implicated with volition in other studies, such as lateral intraparietal neurons—linked to proactive movements (Maimon and Assad, 2006)—and other parietal areas linked to conscious willing (Assal et al., 2007),

were simply not clinically relevant in any of the patients and therefore inaccessible in the current study. Apart from obvious drawbacks such as greater noise, several apply specifically to the LFP analysis. For example, our regional coherence studies were limited to the entirety of channels within that region (rather than, for example, just those channels exhibiting significant deviations in pre-W LFP power at the same band frequency being analyzed for synchrony). Since coherence requires pair-wise interactions from simultaneously channels, pseudopopulations of channels cannot be created (as they can with calculating absolute power); rather, coherence studies are confined to all of the trials for a given patient in a given session. The associated number of significant channels in any given experimental session was too small for coherence analyses, so we relaxed our requirement by considering all pairwise channel interactions, whether or not they had been previously validated as significant. This again introduced noise into coherence calculations, and more refined analyses, such as temporal profiles of coherence over time, examinations of phase in relation to LFP coherence, and even Granger causality, might be possible if the experiment were done with more channels and more trials. Indeed, our current analysis lies silent on the question of causality: given that regions like the SMA proper and pre-SMA are communicating with one another, which one begins the conversation?

The limit of experimental samples also might help explain the performance of the classifer. While predictions of deviation from baseline were accurate at levels higher than those expected by pure chance, our classifier performed worse than similar classifiers tuned to spike rates rather than LFP signatures (Fried et al., 2011). Whereas a pseudopopulation of unit's spike firing signatures could have a sample size in the hundreds, the highest pseudopopulation sample size available for a single given significant predictor, high gamma in the SMA, was 18 in this study. And, as previously stated, since coherence is limited to a given session, we were unable to use as input into the classifer any predictors related to coherence.

Such considerations offer multiple areas for extension in future studies, apart from having larger sample sizes across more regions for further regional interrogation. First, more refined coherence studies should better dissociate the functional and directional relationships between the synchronous regions only identified here. Second, intracranial Libet studies have been only LFPs (ours) or only spikes. It is appealing to conduct a future hybrid study. Recall that spikes and local field potentials are superimposed. Conducting an analysis considering hybrid metrics such as spike-LFP coherence would add another layer of understanding, and, more importantly, possibly elucidate how local field potentials in volition govern or regulate actual spike timing. The conjoined consideration of simultaneous action potentials and the lower frequency waves they "ride on" will almost certainly give higher predictability than either metric alone. In so far as predictability is the most important aspect of Libet studies for bearing on free will (see the next section of the discussion), this may prove critical, particularly in Libetian-styled studies including multiple alternatives (in contrast with this "one-option" task, as will be discussed in the next section).

Nevertheless, experimental pitfalls and necessary extensions notwithstanding, our results mark the first ever attempt to fill in the local field potential story during volition. In so far as displays of neural activity antecedent to conscious volition has any bite in the philosophical problem of free will, our findings are consistent with Libet's original *bereitschaftspotential* with far more frequency decomposition than he could have imagined. In so far as there is bite to the notion that spelling out the when, where, and how of volition in different brain regions will, as Greene and Cohen (2004) argue, makes us less likely to believe in free will, our results, particularly coherence between various frontal lobe regions, supplement the current body of knowledge. And in so far as there is any bite to the threat of predictability for free will, our statistical classifier, while poorer than previous classifiers for different predictor types, performed well above chance. It is whether any of these conceptual premises actually have bite to which we now turn.

Discussion: Can Libet experiments revolutionize the free will debate?

1. Incompatibilist causal histories

If we refresh ourselves with the rough sketch of the free will dialectic presented in the introduction, we see that a hard determinist's *modus ponens* is a libertarian's *modus tollens*. While both agree that free will is incompatible with determinism, no sooner do they step onto common hypothetical ground—if the universe is deterministic, then we lack free will—than does the libertarian, knowing full well that we have free will, negate the consequent and conclude that the universe is not deterministic; and the hard determinist, knowing full well that the universe is deterministic,[2] affirms the antecedent and concludes that we lack free will. Their disagreement lying in a question about the causal history of our decisions, it is no surprise to see neuroscience of the sort explored in this thesis, which pins the brain as the relevant causal locus of choice, creeping in to try to fill out an empirical premise about the etiology of our decisions. Philosophers have just been, so says my optimistic neuroscientist, speculating about for thousands of years what a few clever instances of empirical investigation could settle for good.

He clearly thinks that lab-chair inquiry in ivory-colored buildings is of some importance to arm-chair thinking in ivy-coated buildings, and whether or not this covers much of the free will dialogue, it is worth spelling out exactly what contribution neuroscience is *in the running* to make. What *could* it tell us about free will? I focus here on the relevance of Libet studies toward incompatibilist strands of freedom (particularly agent-causation), arguing they could tell us much less about libertarianism than many neuroscientists think, but probably more than the libertarian would like. I conclude with brief, pessimistic remarks about neuroscience's ability to bear on compatibilism.

2. Libet's paradigm and the classical argument

To rehearse the results described in this thesis and many other Libet studies, the essential

[2] And (if he's a hard incompatibilist) that even if it weren't, such randomness could not possibly procure freedom.

Libet finding is that neural activity, presumably what ultimately leads to some action, is temporally antecedent to the conscious decision. Many interpret this as showing that, despite our experience of deciding, our actions are actually caused nonconsciously; our conscious wills are just us becoming aware of a decision our brain has already made—and with this launched decades of "willusionist" apologetics that have left many philosophers unimpressed ever since.

Whatever justificatory juice is coming from the Libet lemon seems to rely on the fact that nonconscious brain activity is temporally prior to the conscious will. But to see whether or how this bears on free will, it is instructive to spell out the argument that scientists seem to appeal to:

I. Dissociation from conscious decision
1. In Libet-actions, nonconscious brain events reliably precede conscious decisions to act.
2. If nonconscious brain events reliably precede conscious decisions to act, then nonconscious brain events initiate the action.
3. In Libet actions, nonconscious brain events initiate the action. [*follows from 1, 2*]
4. Nonconscious brain events are not identical to or part of conscious decisions.
5. Libet actions are not initiated by conscious decisions. [*follows from 3, 4*]

II. Conscious decisions and free will
6. If an action is an instance of free will, then it is initiated by a conscious decision.
7. Libet actions are not instances of free will. [*follows from 5, 6*]

III. Extrapolating to all actions
8. If Libet-actions are not instances of free will, then no action is an instance of free will.
9. No action is an instance of free will. [*follows from 7, 8*]

The basic thrust is that in so far as our understanding of free will relies on our conscious will to be what initiates action, the preceding neural activity renders our experienced choosing inefficacious in some sense. In the remainder of this thesis, I suggest that 1) the argument presupposes an incompatibilist notion of free will; 2) the argument, as it stands, can be plausibly blocked by the libertarian on conceptual grounds; but 3) the paradigm dovetails predictability considerations that could eventually pressure the libertarian, and briefly 4) Libet results poorly apply to compatibilist accounts of freedom.

3. What does Libet mean by free will?

It is crucial to flag the conceptual analysis of freedom that allows the argument to run:

6. If an action is an instance of free will, then it is initiated by a conscious decision.

Both compatibilists and many libertarians will likely raise their eyebrows at this stipulative claim. This sort of analysis of freedom is almost certainly the very high-grade level of freedom that compatibilists have tried to avoid; and the agent-causal libertarian,[3] in so far as he stipulates that the conscious will is the agent-*cause* in virtue of which our actions are not determined, will probably find it suspicious that the word "initiated," rather than "caused," is used.

I will address the compatibilist's worry in *8-10*, and the libertarian's worry in *7*. For now, I only wish to emphasize that the premise seems to presuppose a highly incompatibilist conception of free will. If the stipulation is that, in order for an agent's actions to be free, the agent's conscious will must be a *causal origin* or an effective source (must be the thing in virtue of which the alternatives are genuinely open) in his decision, that our conscious will must make some *causal contribution over and above* what's encoded in RPs for our act to be free, the concept of "free will" being used is one of a strictly libertarian[4] (and likely agent-causal[5]) variety. Hence, even if the argument is correct, then all that would follow would be a rejection of agent-causal libertarianism.

---

[3] I focus on agent-causal libertarianism here for one primary reason: only agent-causal libertarians (rather than event-causal ones) will very likely want to place the locus of indeterministic causation at the moment that the immaterial consciousness or spirit or soul causes behavior (which seems to be the dualist picture of decision-making that Libet himself originally sought out to challenge), and this is a crucial feature of the libertarian accounts that I will argue neuroscience can challenge (in so far as other libertarian strands of freedom, like event-causation, share similar requirements, which I shall flag, such findings will apply to them as well).

[4] While I find it most immediate to understand Libet's argument as presupposing incompatibilism, this need not be the case. After all, as Kane (2005) states, "If conscious willing is illusory or epiphenomenalism is true, all accounts of free will go down, compatibilist and incompatibilist." The Libet challenge might be one that says that conscious will makes no causal contribution to action, and that any plausible compatibilist account must include conscious will in the causal antecedents of an action. However, if this is argued, then the temporal precedence of nonconscious brain activity becomes less worrying (after all, RPs could just be a causal antecedent to whatever downstream neural activity the mental states of conscious willings supervene upon). In any case, in so far as Libet's results pertain to compatibilist varieties of freedom, I shall deal with those implications in *8-10*.

[5] If we understand the conceptual claim as requiring there to be some extra causal "oomphiness" after the RP, this most naturally will apply to agent-causal accounts of free will which place the agent-cause at the moment of conscious will.

5. Why the experiment falls short

But is the argument correct in the first place? An enormous amount of the philosophical ink has been spilled to suggest that the actual experimental set up does not warrant conclusions drawn from the results. For example, common objections might include worrying:

1. There is a time difference between consciously deciding and consciously accessing the content of decisions. Even if not, subjects systematically lag reporting W-time (Dennet and Kinsbourne, 1992).
2. RPs come from averaging blocks of trials, so reflect nothing about individual trials (Stamm, 1985).
3. RPs are identified from *post hoc* data analysis based on the wrist flex, so could just be fluctuations in cortical activity (Eccles, 1985).

I believe these are all excellent criticisms of the current methodologies, but these are not the sort of challenges I wish to levy—for two reasons: one dialectical and one rhetorical. Dialectically, these sorts of challenges only leave us agnostic to the relevance of neuroscience. They are all challenges which more sophisticated experiments could address.[6] So, even if they are legitimate, these challenges only reveal that neuroscience *cannot now* disprove agent-causation—but it is left open that it *could*. Indeed, the present study makes at least moderate progress on addressing (1) and (2), particularly addressing (1) by constructing temporal profiles of channel recruitment to show significance long before the standard pre-W period, and addressing (2) by using machine learning to discriminate LFP signatures on a single trial basis. But besides this, there seems to be a more important reason: rhetorically, if we challenge the neuroscientific results on the basis of their flawed methodology, it may come across that we are implicitly suggesting that *if* such results followed from more methodologically sound or robust investigation, the conclusions would follow (the philosopher seems to be banking on a flaw in obtaining the empirical results, not with their implications). For both these reasons, I wish to rather explore the conceptual limits of what these

[6] For (1), the further back the temporal correlations go, the less likely that the difference in W-time and the RP are due to systemic bias in reports or reports failing to reflect when the decision actually was intentionally made. For (2), better spatiotemporal precision in recordings could give resolution at the individual trial level. For (3), we do the messy task of seeing cortical activity for the entire duration of the experiment, rather than *post hoc* identifying RPs. Perhaps when we do any of these we would still find the neuroscience to fall short; but it might not—we are simply left agnostic.

sort of Libet experiments could show, given that there is an RP before the conscious intention, and that sample sizes are sufficiently high, and that this mechanism generalizes to all actions, and so on. If one can block the argument on conceptual grounds, no methodological fussing need take place—and I believe that the classical argument, as I've presented it, can be blocked in this way.

6. Why the classical argument falls short

Let's concede any worries about experimental methodology. Does the agent-causal libertarian, in so far as he buys these findings, need to abandon his position?

Probably not. The agent-causalist can begin by pointing out that nowhere in his model is a denial that neural events precede conscious wills. Indeed, the libertarian can accuse Libet of confusing initiation for causation. It may be true that neural events initiate the action, the libertarian says, but this is *wholly consistent* with the agent-cause later jumping in to actually sufficiently determine the choice. My brother *initiates* my choice of chocolate ice cream when he presents me with chocolate or vanilla, but he does not cause it. Thus, the question is really whether the RP is what *causes* the action, because if it did, and it occurred temporally prior to the conscious will, *then* we might start making a case for epiphenomenal conscious wills and therefore inefficacious agent-causes. But the Libet-sympathizer is not entitled to simply replace "initiate" with "cause" at any point in the argument, for several reasons. First, initiation is plausibly arrived at here from temporal precedence, but causation is not. Again, whenever I visit my friend, I may hop in my car such that my driving in my car is perfectly correlated with ringing his doorbell. Perhaps my driving in the car could be said to initiate the ringing of his doorbell (but perhaps not even). Certainly it might be necessary for me to do so (he lives far away). But my driving in the car is not what causes the ringing of his doorbell, simply in virtue of the fact that it reliably temporally precedes my extending of my finger to press the button. Thus, if we only have the fact that RPs always precede the conscious decision, there is no way for the paradigm to distinguish between the RP indicating that there will be

a decision and just being a necessary condition for a decision to be made—it might be *necessary* for a decision to be made, but not *sufficient*. Perhaps the immaterial agent-cause gives this final sufficiency.

But it gets even worse than this, for the Libet paradigm is often set up not as a choice between two alternatives, but as a choice between one option and not choosing. I want to suggest that this is highly relevant to the extent to which Libet findings can generalize. For example, suppose we are certain that RPs are also *sufficient* for wrist flexes (RPs *always and only* precede flexes such that if there is an RP, we're guaranteed a wrist flex). Though RPs now tell us that there will be a decision, we still need to distinguish between that and *which decision is* made—does the RP encode the *content* of the decision, rather than the fact that a decision will be made? *If* we understand the situation that Libet participants are in as one in which they are constantly, each second, forcing themselves to make the decision to flex or not flex, then an RP always and only preceding the wrist flexes would be indicative of encoding content. But this presupposes a certain way in which the participants view the experiment: one in which they are continuously making the choice of whether to flex their wrist or not. I find this implausible. Consult your own phenomenology over the next minute if I tell you to periodically decide to flex your wrist and bear in mind when you make the decision to flex. Rather than thinking, "Make-a-decision→yes, flex" or "Make-a-decision→no, don't flex" I find it far more plausible that participants are simply waiting for urges to come up and then acting on them when they become aware of such urges to flex. If this is in fact what's going on, then subjects are not *constantly making decisions*, but are periodically waiting for an urge to come up, and once it pops up there is *only one possible "decision" to be made:* to act on the urge.[7] Under this more plausible experience of participating in Libet studies, the fact that RPs always and only precede the wrist flex is *consistent* with the RP generally being necessary and sufficient for *some decision being made* but not *which alternative is chosen*; it is just that the subjects have been primed to only have one "alternative" when they

[7] Mele (2004, 2008) has made similar arguments on distinguishing feeling an urge and deciding among alternatives.

decide—to flex. If this is true, then the paradigm of deciding to flex one's wrist will be disanalogous from even deciding which hand to raise. Thus, given even perfect methodologies for the classical Libet paradigms, it seems that the agent-causal libertarian can always claim that RPs are simply energetic ramp ups before the agent-cause kicks in, or at best determine that a choice will be made, but do not specify which choice will be made. Perhaps the neuroscientist finds it metaphysically uneconomical or even downright implausible that the universe is set up this way. But this is an independent objection raised from the arm-chair, not with the lab results.

7. What a good argument might look like

I have just argued that, even in its most methodologically robust iterations, the classical Libet experiments do not constitute a powerful strike against agent-causal libertarianism, yet alone the radial claims against free will that many suppose. However, in recent years, a promising extension to Libet experiments has come from paradigms that give subjects actual alternatives (such as pushing a button with one's left vs. right hand), and revealed scientists' ability to not just find prior neural correlates, but rather to *predict* participants' decisions at some point before W-time. As before, I will ignore methodological deficits and instead explore the conceptual limits of such predictability-styled tests. After all, any evidence of *determinism* would be straightforwardly antithetical to an *indeterministic* agent-cause, and predictability seems to be a close relative of causal determinism in a way that I've argued initiation is not.

Or do I speak too quickly? As Adina Roskies notes:

> "[P]redictability is at best a poor cousin to determinism, and one that can betray its familial roots. Although a deterministic system is in principle predictable, in practice predictability is not a guide to determinism. What appears to be stochastic behavior at one level could be the result of determimistic processes at a lower level" (Roskies, 2010, 112).

As a challenge to neuroscientific attempts that claim to "operationalize our understanding of determinism in terms of predictability," (Roskies, 2014, 105) this is an odd response. Indeed, *lack* of predictability is probably a poor guide to indeterminism, which Roskies rightfully notes could either

be due to true metaphysical indeterminacies, or be consistent with determinism that we simply have poor understanding of. Likewise, the fact that a deterministic system entails, in principle, predictability tells us nothing about predictability being a good guide to determinism. But these claims are orthogonal to the question of whether the *presence* of high predictability of actions would justify us in believing that those actions were part of a deterministic system.[8]

Given the right data, I do think a significant degree of predictability would allow us to justifiably infer that we behave deterministically. But before diving into the meat of this, I want to flag that, in order to demonstrate that we behave deterministically, one need not demonstrate that the universe is a deterministic system. This is hardly a claim for neuroscience. Rather, if we take the quite plausible view that all behavior is orchestrated by the nervous system, and of it, predominantly the brain, then one need only show that *the brain's activity* is deterministic in order to show that humans behave deterministically. If the brain operates deterministically but, in fact, quantum events are genuinely indeterministic, the universe would be indeterministic but these indeterminacies would be irrelevant to human choice. So, the justificatory demand on the Libet-sympathizer is to take us from some degree of predictability in our actions to brain determinism. As a starting point, consider:

(1) If an action is indeterministic, then it is not theoretically perfectly predictable by humans.[9]

If John behaves indeterministically, then some prior state of the universe in conjunction with the laws of nature are not causally sufficient for his action, and are in fact consistent with either of his apparent alternatives being selected in the actual world. But since knowledge of some past state of the world and knowledge of natural laws are all we could use to make our predictions, they will never be theoretically perfectly predictable, even with infinite computing power. But the contrapositive of (1) is just:

[8] Roskies does suggest that one system having higher predictability than another does not suggest that we should have greater confidence that it is deterministic, since indeterministic systems could be predicted more accurately than deterministic systems riddled with chaos. I agree with this, but I make no such claims about relative predictability.
[9] The "by humans" addendum is a move to avoid concerns about divine foreknowledge.

(2) If an action is theoretically perfectly predictable by humans, then it is deterministic.

So rather providing direct evidence that the universe or our brains are deterministic, it will suffice for my neuroscientist to show that Libet actions are theoretically perfectly predictable.[10]

I contend that he could plausibly infer that antecedent of (2)—that Libet actions are theoretically perfectly predictable—if his predictability algorithms become accurate enough. To see why, I would first like to flag three points. First, in all of these Libet-tasks so far, the participant is in some Buridan's ass type scenario where there is no reason for choosing one alternative versus the other. The agent's reasons for either alternative (what I call the reasons-split) are about 50-50. Second, the accuracy of our predictions that we'd expect simply by chance for two-alternative situations is 50%. Finally, we will almost certainly never get 100% predictability of Libet decisions, simply because of how complex brain activity is relative to our instruments for recording it. The implications of this final point are as follows: given that neuroscience will never give us perfect predictability, since presumably nothing is going to *deductively entail* it, the way we could plausibly fill in the antecedent of (2) is through an *abductive* argument[11]: theoretically perfect predictability is the inference to the best explanation (IBE) of some empirical finding. But which empirical finding?

I now want to explore what happens (and what inferences we're justified in making) as the accuracy of our predictions (which has a baseline lower bound of 50%) increases for agent decisions with a 50-50 reasons split. If predictability is only marginally higher, say 60-70%, as has been found

[10] The reader may note that, in allowing neuroscience to show brain determinism, this will not only be an attack on agent-causation; presumably it will attack any *incompatibilist* theory of free will. This is true in so far as the non agent-causal libertarian (say, the event-causal libertarian) places the freedom-enhancing indeterminacy *after* the timepoint of the idealized prediction. If that is the case (for example, if the event-causalist posits quantum indeterminacies at the moment of conscious willing), my argument will apply to his theory as well. However, some event-causalists are indeterminacy historicists: they are not time-slice/Valerian libertarians, and so place the indeterminacy far before any prediction would be made. In so far as an event-causal account postulates an indeterminacy that is prior to time at which neuroscientists make their prediction, idealized predictability would not be evidence of determinism. Of course, if we find that most actions in the lab are theoretically perfectly predictable, the burden on the event-causalist would be to explain the disanalogy between deterministic lab actions and the ones which he wishes to call free, but I don't wish to dive seriously into this part of the dialogue.

[11] The type of abduction I will be using will be inference to the best explanation (IBE).

in current studies, then the volitional implications of the results succumb to a widespread criticism of these predictability studies in the philosophy literature: a far better explanation is that the predictability reflects (non-agential) *biasing* that predisposes the agent to one alternative (non-conscious *inclinations*) but does not fully determine the decision[12] (Levy, 2014, 25-25). In this case, we don't infer that the gap between 60% and 100% predictability is due *solely* to epistemic deficits of our recording instruments and algorithms, but rather that the action is not theoretically perfectly predictable in the first place—the relevant deviation is the agent-cause determining the outcome, albeit biased from 50% to 60% (toward one alternative). If true, this would only reveal what might be called nonconscious neural influencers of behavior that increase the probability of action, but are not causally sufficient for it. This would make for poor fodder against agent-causation, since indeterminate agent-causes could be fully compatible with non-agential biases, prior inclinations and proclivities that do not causally determine the decision.

But perhaps this is too uncharitable a reading; the neuroscientists are probably excited because they expect their predictions to become more accurate as their technology becomes more sophisticated. This does not seem too unreasonable, so let us suppose the accuracy of predictions continues to rise. If the accuracy rises, from 60% to 80%, to perhaps 95%, this starts tipping the IBE in favor of theoretical perfect predictability rather than mere biasing. What other competing explanations would explain the high degree of predictability? Surely chance is a poor explanation. Moreover, the other competing explanation would be that our agent-causes are extremely biased by underlying physical activity. But an increasing accuracy of predictions would correspond to an ever decreasingly efficacious power of the agent-cause (to generate metaphysical alternatives), and under this trajectory, it seems that the weaker the agent-cause gets the less plausible that it exists at all. I

[12] The example, in the case of lifting one's left vs. right hands, would be that one still agent-causes which hand is lifted, but some prior neural activity inclines one a certain way. The nonconscious brain activity would certainly be a *determinant* of behavior, but only in the same way as smoking is a determinant of cancer. And a biased agent-cause could still be an indeterministically behaving agent-cause.

will not deal here with what percentage would justify the inferential transition from simple influencers of action to idealized predictability (and so full blooded determinism), or the requisite sample sizes—these substantive questions exceed the scope of this thesis. I merely claim that, at *some* point, the neuroscientist is entitled to infer that Libet actions are theoretically perfectly predictable, and therefore the neural events underlying them are deterministic.[13]

At this point it seems like the agent-causalist can quite easily push my neuroscientist into a nasty dilemma. Recall from earlier my challenge that Libet experiments only deal with Buridan's ass type scenarios. In predictability contexts, a potential disanalogy between reasonless and really difficult decisions may elevate to an incredibly worrying conceptual challenge. Why? Recall that the neuroscientist is only entitled to his theoretical perfect predictability IBE *not* when accuracy is high, but when it is *much higher than the reasons-split for the behavior*. The reason something like 95% accuracy might merit an idealized predictability IBE is because the comparison was to 50-50 reasons-split, and this was because the agent was in a Buridan's ass scenario. If we *were* to start doing Libet studies with more valenced decisions or ones in which the agent had prior reasons, simply getting high predictability would not necessarily warrant the IBE. If I love pistachio ice cream and hate pink toe jam, and you know this, you can probably predict which alternative I will choose (without even measuring my brain!) if I'm given the choice. So the agent-causalist can plausibly defend that brain activity prior to the decision is just reflecting the agent's prior reasons (it reflect an agent's conscious reasons and inclinations rather than just nonconscious physical biases); so it doesn't matter that you can predict with 100% accuracy my choosing the pistachio ice cream—I still agent-causally take my frozen dessert. So *either* the neuroscientist only conducts Libet experiments for Buridan's ass type

[13] Since this is an abductive argument, presumably the introduction of new information might make postulating indeterminism a better inference than idealized predictability. What might this look like? Suppose that in fact we are agent-causes, but the biasing spells out the highest possible predictability of our actions (since our actions are to some extent indeterministic and so inpredictable). As technology improves, our predictions are of increasing accuracy until they plateau at a certain value and never go higher. For example, if no matter how much how technology improved, our predictions could never get better than 70% (and we had near perfect resolution of the brain in making these predictions), then idealized predictability is not so plausibly the best explanation of the accuracy. It may just be biasing.

decisions, in which case there will always be a potential disanalogy between the deterministic nature of Buridan's ass type decisions made in Libet experiments and the libertarian free actions made in every day life; *or* the neuroscientist will include experiments where the agent has reasons that might make him lean toward one alternative or the other, or simply reasons that seem equal or incommensurable and make the decision very difficult, and in that case any predictability could just reflect the brain activity associated with or generated by prior (nonphysical) agential reasons, desires, and inclinations.

But my neuroscientist seems quite able to grab the second horn of the dilemma. All that's needed is to remember a large motivation for agent-causation in the first place: one is pessimistic that the mechanistic explanations of science and the purposive explanations of tying actions to beliefs and intentions are compatible, and since purposive explanations are not illusory (the phenomenology of action is truth-tracking), there must be some fundamental or irreducible or primitive notion of an agent. The non-illusory nature of phenomenology is crucial—we experience our alternatives as live options[14], and believe that through conscious will we can resolve the decision. So, when the agent is making a decision, we no longer have the easy assumption of a 50-50 split (because the agent has no reason for either alternative), but things *must seem a certain way to the agent phenomenologically*; he may see this as an extremely torn decision where he regards it as a 50-50 split, or perhaps is leaning toward one alternative. Whatever it is, the experimentally given reasons-split is now replaced with whatever the agent *regards*[15] as his reasons-split, and so long as my neuroscientist can predict the decision with far greater accuracy than the agent's phenomenology of his reasons (the agent's subjectively identified reasons-split), then he can plausibly make the IBE that the decision was theoretically perfectly predictably (even when the agent considered it not to be). In this

[14] I actually believe it's an open question whether we phenomenologically experience libertarian or contracausal freedom. At best, I experience being able to choose an alternative if I were in a qualitatively identical state. But since we are only conscious of a subset of the events that determine our decision, this is consistent with determinism.
[15] We may want to insert idealized conditions, such as: what he *would regard* if he fully thought about his prior reasons.

case, the agent's phenomenology would not be truth-tracking—and so the predictive neural activity is *not* simply a measure of his prior nonphysical reasons. Suppose that, after we did all the philosophy of science, we decided that for insignificant no-reason split actions like button presses, with 50-50 experimentally given splits, 95% predictive accuracy would be enough to IBE idealized predictability. If participants experience making extremely difficult decisions where they seem to have equally strong (but different) reasons for each alternative, then it seems like 95% predictability would also warrant the same IBE. If they experienced leaning toward one alternative, the requisite percentage for predictable accuracy might be higher, but the explanation that prior, nonphysical reasons can manifest in predictive neural activity seems to get trumped when the neuroscientist can predict the participant's decisions far more accurately than the subject takes his options to be live.

It is worth noting that even if the agent-causalist does not find any predictive accuracy short of 100% to merit an IBE to idealized predictability, the neuroscience still complicates agent-causation. Even if extremely high predictive accuracy show the extent to which our actions are biased, but not determined, and there is still some elbow-room for the agent-cause to jump in, this is markedly weaker than what most agent-causation theories make the agent-cause out to be; it would mean that *biasing* of the ultimate decision is extremely large. Neuroscience is far from anywhere near the kind of predictability discussed in this section, but it seems possible that it should arrive there—and if it did, this would weaken the agent-cause one way or another.

8. Where did compatibilism go?

Because of an impending page limit, I must close with all too cursory remarks on Libet experiments in relation to the compatibility question—the question of whether free will is compatible with determinism. After all, I have only argued that Libet results could, *in theory*, cast pressure on some forms of libertarianism. But if neuroscience is supposed to revolutionize the free

will debate (by disproving it, or something along those lines), then it must deal with the other main variety of free will: compatibilist freedom. Can Libet sink the compatibilists with the libertarians?

To sketch out exactly what neuroscience must do to settle the compatibility question, let's flag two questions that seem to nicely summarize what we need answers to:

Metaphysical question: What kinds of freedom do we have? (What is the causal nature of human-decision making?)
Conceptual question: What does it mean to have free will?

It seems like if we know what abilities we have, and what are needed to have free will, then we will be able to answer whether we have it. The upshot of this divide is that we can see that the first question is straightforwardly a question about the metaphysics (or perhaps, physics) of how we choose (i.e. are we agent-causes with contracausal powers?), and presumably the one that neuroscience tries to gain traction over. On the other hand, the conceptual question of what it actually takes to have free will (i.e. do we need more than what determinism provides?) *prima facie* seems much less tractable by science. But since we need answers to *both* questions to settle free will, *no empirical finding about our decision capacities can generate a conclusion about free will without fixing a conceptual understanding of what it takes to have free will.* So to disprove free will, neuroscience must somehow provide some evidence of what the right concept of free will is, and that means settling whether free will is incompatible with determinism. Can it do that?

To start, let us concede that my sympathies to neuroscience in the previous sections have been on point: Libet experiments can theoretically provide evidence of neural determinism, and in this way provide evidence against agent-causation. But simply providing evidence of (neural) determinism is the very *starting* point of the compatibility question, so that clearly won't do. Indeed, there seems to be only two possible ways in which some sort of neuroscience finding from Libet-styled experiments could inform the compatibility question: either 1) they show that a specific sort of form of determinism underlies our actions, and it's obvious that that specific form is freedom-

undermining in a way that general determinism is simply vague about; or 2) they justifiably change our judgments about the compatibility question. I consider and dismiss each of these in turn.

9. Take 1: Neuroscience reveals a particularly freedom-undermining variety of determinism

Maybe determinism *simpliciter* does not threaten free will, but—as the first argument goes—everyone agrees that P-determinism (a specific sort of determinism) would undermine free will, and neuroscience is shows that we live in a P-deterministic universe. For example, suppose that when we started looking into people's brains, we found, *per impossibile*, that all of our actions are actually determined by some mindless (deterministically acting) demon who then zaps the relevant parts of our brains and creates an illusory sense of ownership to go with it. Then neuroscience might play a justificatory role because it shows that we're governed by a *type* of determinism that is imcompatible with freedom. The problem here is that there is no chance that neural determinism fills the role of this P-determinism. Compatibilists argue that free will just means being responsive to reasons, or acting in accordance with character, or [insert your favorite compatibilist analysis of freedom], and none of these claims to work with some form of determinism but *not* with neural determinism.[16,17]

Of course, the neuroscientist is more than welcome to retort with something along the lines of: you agree that some demon externally determining all of our actions would be freedom undermining, but there is no difference relevant to freedom or moral responsibility between that case and one in which our actions (or what we internally decide to do) are nevertheless externally determined by some arbitrary prior state of the universe in conjunction with the laws of nature. And maybe this is a good analogy, and constitutes a good strike against compatibilism, but that is an objection takes place in the arm-chair—the neuroscience results have dropped out of the picture.

---

[16] I have never seen, nor can I imagine anyone arguing, that we are free so long as we act in response to our reasons, unless the causal history of our responsiveness to reasons is through action potentials and LFPs…

[17] In fact, for most compatibilist versions of freedom, it is almost *obvious* that we possess it; it is obvious that we are sometimes responsive to reasons or that we have wills structured in such a way that we often endorse the action we ultimately will. The contention with compatibilist accounts of freedom is not whether we possess them, but whether they are *worth* calling freedom.

10. Take 2: Neuroscience directly changes the judgments we form about compatibilism

Given that the first attempt is unlikely to ever pan out, we now turn to a far more plausible (though as I shall argue, unsuccessful) way in which one might argue that neuroscience can tackle the compatibility question: neuroscience can better acquaint us with the concept of determinism, or more vividly show us what determinism looks like, and this new information makes it more clear how determinism and free will are incompatible. Such seems to be the sort of argument offered by Greene and Cohen (2004, 1781)[18]:

> As long as the mind remains a black box, there will always be a donkey on which to pin dualist and libertarian intuitions. For a long time, philosophical arguments have persuaded some people that human action has purely mechanical causes, but not everyone cares for philosophical arguments. Arguments are nice, but physical demonstrations are far more compelling. What neuroscience does, and will continue to do at an accelerated pace, is elucidate the 'when', 'where' and 'how' of the mechanical processes that cause behaviour…At some further point [brain technology that can completely map and predict decision making] may be very widespread, with a high-resolution brain scanner in every classroom. People may grow up completely used to the idea that every decision is a thoroughly mechanical process, the outcome of which is completely determined by the results of prior mechanical processes. What will such people think as they sit in their jury boxes?...We submit that these questions [about retribution], which seem so important today, will lose their grip in an age when the mechanical nature of human decision-making is fully appreciated.

I find it extremely difficult to understand exactly what argument is being made about neuroscience and compatibilism—there seem to be three possibilities. I consider each in turn.

First, the authors might be suggesting what they later call a "transparency bottleneck," with the basic thrust being that all of our intuitions about free will, moral responsibility, and retribution are straightforwardly libertarian in nature, and they creep in when we don't see the full deterministic picture (as long as we are unaware of the causal antecedents to our actions, we assume there is some wiggle room for an agent-cause). Thus, once neuroscience vividly demonstrates that the whole shebang is determinism, we drop those judgments of freedom. This clearly begs the question against compatibilism, since it assumes that once we fully understand determinism (through neuroscience), there will be no argument as to whether we are free or not.

[18] This paper is written specifically against retribution (or the freedom needed to justify retribution), but the form of the argument could easily be adapted for moral responsibility or free will in general. I am less interested in whether the authors actually are making this argument; I am more interested in the form of the argument that might be appealed to.

Second, the argument might be that neuroscience will help elucidate what determinism means/looks like, and in doing so change people's judgments about the compatibility of free will with it. At least the elucidation part is correct; no doubt, supposing that technology tracks the path the Greene and Cohen believe, seeing the pathways and connections from stimulus to behavior along with algorithms predicting the entire time course, all alongside glowing holographic neurons will certainly give us a sense of what a billiard-balls-in-the-void model of determinism looks like. People might in fact (and probably will) change their mind about compatibilism as a result of seeing this, but this does not come close to suggesting that neuroscience is playing a justificatory role. Suppose you want to demonstrate to someone that 1+1=2. You show him some incredibly complex abstract proof for why 1+1=2, but the abstract symbols fly over his head. You then simplify things by making them concrete. You show him one block and he agrees that it represents "1" and another block that he agrees represents a "1" and then you press the two together and he agrees that the conglomeration is "2". You have vividly demonstrated that 1+1=2, but your demonstration did nothing justificatory for the math. It was only explanatory.

Now, one reply from the neuroscience-sympathizer might be to suggest that such standards for justification are too high; even if something only explicates a concept, or better acquaints one with it, it is still playing a justificatory role—and that is because thought experiments function exactly like that. After all, my neuroscientist reasons, suppose you learn about act utilitarianism in an ethics course and agree from the get go that the optimific thing is obligatory. But then a few days later you imagine that this would require you to kill an innocent pizza delivery boy and harvest his organs to save five dying patients who need transplants, and decide that this is obviously a good counterexample, and so conclude that act-utlitarianism is false. The thought experiment definitely seems to play a justificatory role, and it shouldn't make a difference whether you imagine the pizza boy, or happen to walk by your local hospital, see a doctor who has spent too much time reading

Sidgwick kill the delivery boy, and then reject act-utilitarianism when he justifies himself. So too with free will; in so far as thinking about what deterministic decision-making entails leads one to reject compatibilism, it shouldn't matter whether he imagines determinism or sees it through a brain scan.

But it's important here to be mindful of an -ing/-ed distinction. When my neuroscientist says that the thought experiment plays a justificatory role, this is ambiguous between the *action* of thinking of the pizza boy or determinism, and the *content* of the thought (the proposition being entertained). So too with actual perception; distinguish the action of seeing the pizza boy get murdered/the brain scan showing neural determinism with the content of those perceptions (i.e. the proposition that a pizza boy is killed being consistent with act utilitarianism). In all of these cases, it is the content—the imagin*ed* or the perceiv*ed*—that is relevant to justification, not any act of imagin*ing* or perceiv*ing*; if my friend tells me about a thought experiment that changes my mind about some philosophical issue, it is the content of what is told that justifies a change in belief, not his telling. This is the slip in the analogy between thought experiments and neuroscience: if the neuroscientist says thought experiments play a justificatory role, he must be referring to their content; by contrast, the neuroscience, *qua* neuroscience, only constitutes the perceiving rather than the content perceived (which is the concept of determinism), and the perceiving is not playing any justificatory role. So again, all neuroscience is doing, by vividly painting determinism, is helping one better understand what determinism means, but this alone does nothing to justify any move to incompatibilism that might follow; in so far as one's belief changes because they better understand determinism, neuroscience has done nothing *qua* neuroscience any more than my friend telling me about the pizza boy counterexample has justified my drop in act utilitarianism *qua* his telling.

I do not deny that there may be cases in which the perceiv*ing* can play some justificatory role, but only if there is something peculiar about such an experience that goes over and above simply understanding the propositional content of what is being perceived. Highly appealing candidates

seem to be cases like when we are forming judgments about the moral duties we have toward suffering.[19] For example, suppose you are considering the demands of ethics, and the extent to which you should donate your income to those suffering from extreme hunger in developing countries. I find it extremely plausible that if you, for some reason, experience extreme hunger for the first time, and your judgment about your moral duties changes as a result, the experiencing will have played some justificatory role—there is something about the concept of hunger that experiencing it gives special epistemic access to—you now know *what it's like*.

This segues into the final argument that Greene and Cohen might be making. In order for neuroscience to actually play a justificatory role—in order for it to actually bear on the compatibility question—it must do something *over and above* simply acquainting someone with the concept of determinism; there must be something about the *perceiving* that goes over and above just content of what is perceived. But does watching a brain scan give special access to what determinism really means in a way that experiencing hunger does? I find this claim incredible; there is certainly something that it's like to act deterministically[20], but knowing what it's like hardly seems relevant to the judgment of whether determinism is compatible with free will. Moreover, even if it were relevant—suppose our universe is actually deterministic, and, *per impossibile*, if we could experience contra-causal agency we would immediately become incompatibilists—this would not be something that we could understand by *watching* a brain scan.[21] If there are relevant factors contained in a brain scan that explicate determinism in a way not possible from the armchair, in a way that goes over and above just thinking about what it means to live in a clockwork universe, I am not aware of them.

11. The limits of Libet

If what I have said is true, neuroscience will at best adjudicate *within* incompatibilist

[19] I am indebted to Susanna Rinard for discussions helping formulate this.
[20] We might even experience this, if in fact our universe is deterministic.
[21] Given that we either do or don't exist with such capacities, it's not something we could ever fail to experience or experience, respectively

positions, casting doubt upon agent-causal varieties that place the agent-cause at the moment of conscious will, and upon event-causal varieties that locate the freedom-enhancing indeterminacy at any point after the high accuracy predictions are made. But such incompatibilist accounts of freedom make up only a tiny portion of the dialectic—about 10% of philosophers, by some surveys (Bourget and Chalmers, 2013)—which dominantly centers around the compatibility question, a question which neuroscience bears little on. It seems that a fairly good heuristic is that the only varieties of freedom worth disproving from a lab-chair are those worth wanting from an arm-chair. If we can, from the arm-chair, decide that whatever free will experiments are in the running of casting doubt upon are either incoherent or metaphysically implausible (as many argue about agent-causal libertarianism), our resources are probably better directed toward running *thought* experiments.

Now, this is not in any way to dismiss the importance of Libet experiments. While very few philosophers seem to be libertarians, if it is true that our legal system's current punitive sentences would only be justifiable if we are all full-fledged agent-causes with contracausal freedom, and I suspect this is the case, then these experiments may play a pivotal role in dissuading the wider public, which does not have the luxury of extended arm-chair theorizing about volition, from what may be fundamentally incompatibilist intuitions. And if our tendencies to hold grudges against one another, or treat pedophiles as worse than murderers, or condone the suffering of those we hate, if any of these things also rely on strictly incompatibilist conceptions of ourselves and of each other, then perhaps Libet experiments can bring about a great deal of positive change. But potential for social progress does not generate philosophical juiciness, so the importance of such social change notwithstanding (even if, as I suspect, such progress would be more valuable than actually solving the problem of free will), we ought to be clear about what philosophical implications these experiments actually have, and whether they could eventually solve free will. And neither can Libet experiments currently do this, nor, it seems, could they ever have the ability to do otherwise.

References:

Adesnik H, Scanziani M (2010) Lateral competition for cortical space by layer-specific horizontal circuits. Nature 464:1155-1160.

Ahmed O, Mehta M (2012) Running Speed Alters the Frequency of Hippocampal Gamma Oscillations. Journal of Neuroscience 32:7373-7383.

Assal F, Schwartz S, Vuilleumier P (2007) Moving with or without will: functional neural correlates of alien hand syndrome. Annals of Neurology 62:301-306.

Bansal A, Madhavan R, Agam Y, Golby A, Madsen J, Kreiman G (2014) Neural Dynamics Underlying Target Detection in the Human Brain. Journal of Neuroscience 34:3042-3055.

Baumeister R, Masicampo E, DeWall C (2009) Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness. Personality and Social Psychology Bulletin 35:260-268.

Bourget D, Chalmers D (2013) What do philosophers believe?. Philos Stud 170:465-500.

Brass M, Haggard P (2007) To Do or Not to Do: The Neural Signature of Self-Control. Journal of Neuroscience 27:9141-9145.

Buzsaki G (2004) Neuronal Oscillations in Cortical Networks. Science 304:1926-1929.

Cardin J, Carlén M, Meletis K, Knoblich U, Zhang F, Deisseroth K, Tsai L, Moore C (2009) Driving fast-spiking cells induces gamma rhythm and controls sensory responses. Nature 459:663-667.

Dennett D, Kinsbourne M (1992) Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain Sciences 15:183-201.

Eccles J (1985) Mental summation: The timing of voluntary intentions by cortical activity. Behavioral and Brain Sciences 8:542.

Engel A, Moll C, Fried I, Ojemann G (2005) Invasive recordings from the human brain: clinical insights and beyond. Nature Reviews Neuroscience 6:35-47.

Fried I, Mukamel R, Kreiman G (2011) Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition. Neuron 69:548-562.

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society B: Biological Sciences 359:1775-1785.

Hsu C, Chang C, Lin C (2003) A Practical Guide to Support Vector Classification, 1st ed. Taipei: Department of Computer Science National Taiwan University. Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed February 26, 2016].

Hung C, Kreiman G, Poggio T, DiCarlo J (2005) Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science 310:863-866.

Jarvis M, Mitra P (2001) Sampling Properties of the Spectrum and Coherency of Sequences of Action Potentials. Neural Computation 13:717-749.

Joliot M, Ribary U, Llinas R (1994) Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. Proceedings of the National Academy of Sciences 91:11748-11751.

Kane R (2005) Remarks on the Psychology of Free Will.

Laplane D, Talairach J, Meininger V, Bancaud J, Orgogozo J (1977) Clinical consequences of corticectomies involving the supplementary motor area in man. Journal of the Neurological Sciences 34:301-314.

Levy N (2014) Consciousness and moral responsibility. Oxford: Oxford University Press.

Libet B, Gleason C, Wrist E, Pearl D (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). Brain 106:623-642.

Loewy A (2009) Criminal law in a nutshell. St. Paul, MN: West.

Maimon G, Assad J (2006) A cognitive signal for the proactive timing of action in macaque LIP. Nature Neuroscience 9:948-955.

Mele A (2004) The Illusion of Conscious Will and the Causation of Intentional Actions. Philosophical Topics 32:193-213.

Mele A (2006) Free will and luck. Oxford: Oxford University Press.

Mele A (2008) Recent Work on Free Will and Science. American Philosophical Quarterly 45:107-130.

Mele A (2009) Effective intentions. Oxford: Oxford University Press.

Mitzdorf U (1985) Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. Physiological Review 65:37-100.

Nahmias E, Morris S, Nadelhoffer T, Turner J (2005) Surveying Freedom: Folk Intuitions about free will and moral responsibility. Philosophical Psychology 18:561-584.

Nichols S, Knobe J (2007) Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. Nous 41:663-685.

Renshaw B, Forbes A, Morrison B (1940) Activity of isocortex and hippocampus: electrical studies with microelectro. Journal of neurophysiology 3:74-105.

Roskies A (2010) How Does Neuroscience Affect Our Conception of Volition?. Annu Rev Neurosci 33:109-130.

Roskies A (2010) Why Libet's Studies Don't Pose a Threat to Free Will. In: Conscious Will and Responsibility: A Tribute to Bejamin Libet, 1st ed. (Sinnott-Armstrong W, Nadel L, ed). New York: Oxford University Press.

Schel M, Scheres A, Crone E (2014) New perspectives on self-control development: Highlighting the role of intentional inhibition. Neuropsychologia 65:236-246.

Soon C, Brass M, Heinze H, Haynes J (2008) Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11:543-545.

Stamm J (1985) The uncertainty principle in psychology. Behavioral and Brain Sciences 8:553.

Vohs K, Schooler J (2008) The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. Psychological Science 19:49-54.