THE UNIVERSITY OF CHICAGO

SENSE OF AGENCY ACROSS SCALES AND IMPLICATIONS FOR SELF-AWARENESS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

BY
JOHN P. VEILLETTE

CHICAGO, ILLINOIS
JUNE 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

At the time of writing, the following chapters appear elsewhere as standalone publications. Chapter 1 is in-press at *The Journal of Neuroscience* (Veillette et al., 2025b). Chapter 2 has also been published at *The Journal of Neuroscience* (Veillette et al., 2023b). Chapter 3 has been published at *Consciousness and Cognition* (Veillette et al., 2024). I am grateful to those journals for allowing me to include those works in this dissertation. Even more so, I am especially grateful to my collaborators who were co-authors of these works, including Jacob Rosen, Letitia Ho, Pedro Lopes, and Howard Nusbaum.

I thank the United States National Science Foundation, which supported this work through my graduate research fellowship (DGE 1746045) and through two research grants: BCS 1835181 to Daniel Margoliash and Howard Nusbaum (Chapter 1), and BCS 2024923 to Howard Nusbaum and Pedro Lopes (Chapters 2-3).

I thank my committee members for their contributions and advice, both related and unrelated to this work, throughout my time in this doctoral program: Howard Nusbaum, Pedro Lopes, Daniel Margoliash, and Leslie Kay. I am also grateful to all others who helped me (even in small ways) complete this work, but there are too many of you to name.

I also thank those who helped me procrastinate from working on this manuscript when I needed a break from the workday: Anita Restrepo Lachman, Zeynep Aslan Sisman, Nakwon Rim, Sabina Raja, Elizabeth Gaillard, Claire Guang, Louisa Belian, the *New York Times* puzzle team, and Henry Jones's obscure *PyMC* error messages.

I thank those who gifted me the beer I drank while writing this dissertation, including Anna Corriveau, Alfred Chao, Madeline Sullivan, Jean Boulware, and Shannon Heald.

And I thank my cat, Talisker, who is a delight.

ABSTRACT

To produce effective behavior, it is critical that an organism be able to identify which parts of its sensory world are caused by its actions. On one hand, successful assignment of sensory outcomes to specific actions is critical for sensorimotor learning. Any organism that learns motor behaviors from sensory feedback must have a biological solution to the "hindsight credit assignment" problem, assigning temporally distal outcomes to self-generated actions. On the other hand, we also have a salient conscious experience of causing events in the world through our actions – that is, a sense of agency. The degree to which conscious agency is derived from the same neurocognitive operations that support sensorimotor control has been the subject of extensive debate, with far-reaching implications from the philosophy of self-awareness to motor rehabilitation and embodied interface design. This work examines the relationship between inferences of self-causation made at different scales of cognition. Chapter 1 illustrates the nontriviality of low-level hindsight credit assignment in the context of human speech production, where the nonlinear physics of the vocomotor periphery decouples the timings of acoustic events from that of articulatory movements. Moreover, it presents electrophysiological evidence for a mechanism to restore temporal coherence between neural activity and sensory feedback, the implications of which are elaborated by parallels to findings from the birdsong system. Chapter 2 presents an experiment in which the timing of electrical muscle stimulation is manipulated, using a novel adaptive paradigm, to reliably elicit self-reported agency over externally-actuated muscle movements. Claims of self-agency over externally-caused movements are predictable from the early sensorineural response to muscle stimulation, firmly rooting the origins of bodily sense of agency in sensorimotor processes while also reinforcing a dissociation between conscious and objective agency. Chapter 3 examines the relationship between individual differences in sensorimotor-level control judgments and high-level declarative beliefs about agency. A relationship is confirmed but found to be mediated by conscious introspection (i.e. metacognition), suggesting a preferential

role of consciously accessible mental contents in the formation of self-agency beliefs. Taken together, findings suggest that aspects of low-level movement control impact the sense of agency, even at levels of behavioral organization which are not explicitly motor in nature. The dissertation closes with a critical Discussion of the theoretical assumptions underlying current approaches to cognitive neuroscience. Specifically, I argue that to accomplish the goal of identifying those motor and cognitive processes that impinge upon conscious agency, we must dispense with the categorical error of conflating (especially macroscopic) neural dynamics with direct encodings of mental representations.

INTRODUCTION

Compared to the complexity of the biological machinery required to fluently move a body with roughly 250 joints and over 600 muscles, the conscious experience of movement is quite simple. Despite this mismatch between the objective complexity and subjective simplicity of directing movement, volitional actions are nonetheless accompanied by a subjective experience of control; people (and perhaps other organisms) have a sense of agency, a feeling that "I did that." Subjective experiences of agency are a core component of human self-awareness. Indeed, anomalous experiences of agency in schizophrenia (Frith, 2012), alien hand syndrome (Panikkath et al., 2014), perceived control or non-control of a phantom limb (Ramachandran and Hirstein, 1998), automatic "utilization behavior" in patients with frontal lobe damage (Lhermitte et al., 1986), learned paralysis during stroke recovery (Wolf et al., 1989), and other movement disorders can feel deeply distressing to affected people. In healthy populations, however, the conscious window into motor control processes provided by the sense of agency plays a crucial role in guiding behavior. While it may be distressing to feel of loss of agency when one's car loses traction while driving, this salient experience triggers a high-level motivational program that prevents more dire consequences down the line.

Increasingly, we are seeing technologies inserted directly into our action-outcome loops, such as avatars that replace visual feedback from our bodies in virtual reality, large language models that co-write text with us, and brain-machine interfaces that produce speech or move prosthetics for paralyzed people. Despite the obvious potential of such technologies, they also pose a clear challenge: just as we prefer to consciously feel that our agency has been breached when driving a car, we would also want to consciously detect when speech neuroprostheses – the most successful of which, at present, use large language models to "fill in" information that could not be decoded directly from the brain (Metzger et al., 2023; Tang et al., 2023) – say something we did not intend. While this seems as if it should be trivial, it empirically is not. Manipulations of auditory feedback such that participants hear themselves saying

a word that differs from what they actually uttered often go unnoticed; more unsettlingly, when asked to repeat themselves, participants may repeat the word they heard rather than the one they actually said (Lind et al., 2014). Relatedly, while manipulating the fundamental frequency of auditory feedback usually elicits an automatic compensation response to restore the original pitch (Elman, 1981) – although following, rather than compensatory, responses are also possible (Behroozmand et al., 2012) – manipulations that mimic a vocal expression of emotion alter speakers' actual emotional state in a congruent direction (Aucouturier et al., 2016). In other words, there is a very real risk that, rather than consciously detect deviations from intended actions, our original intentions could be *overwritten* by semi-autonomous technologies. Some governments have already passed legislation to preserve human autonomy in the era of brain-machine interfaces (Fernández and Fernández, 2022). I argue, however, that the enforceability of any such regulation critically depends on the ability of users to consciously detect and report violations of their agency and, in turn, on a basic scientific understanding of how consciousness relates to the sensorimotor (or other action-outcome generating) processes in which such technologies intervene.

Failures to detect sensory feedback manipulations like those described above have been framed as sensorimotor extensions of "choice blindness" phenomenon reported in the social psychology literature, where participants fail to notice when an item they have chosen has been replaced with a non-chosen option (Johansson et al., 2005; Hall et al., 2010), and have similarly been used to argue that the sense of agency and, indeed, intentions themselves are confabulatory, post-hoc constructions. In this view, the sense of agency is not a genuine conscious window into the neurocognitive representations underlying the control of action but an entirely unrelated process. Such arguments, often levied to dismiss the subjective experience of mental causation, are subject to the same weaknesses as are similar views espoused in social psychology (Nisbett and Wilson, 1977; Wegner, 2017). Specifically, the fact that subjective agency seems to be susceptible to post-hoc influences in some settings does not

imply such influences are the sole determinants of agency judgments (Moore and Haggard, 2006). Moreover, intention itself is not a monolith; philosopher John Searle distinguished between "prior intention," or something one aims to accomplish, and "intention-in-action" that is present during any volitional action but would not be considered prior to action initiation (Searle, 1983). As I wrote this sentence, for example, my prior intention was to express a once non-verbal thought, but doing so required many subsidiary actions (e.g. selecting words, individual keypresses) that I had not specifically considered ahead of time but are nonetheless intentional. At a lower level, my nervous system had to complete many subsidiary steps to translate my intentions-in-action into specific neuromuscular instructions; these subroutines too are certainly not unintended, even if they are not necessarily consciously accessible. So, when asked whether one had agency over an action or sequence of actions, to which sort of intention should they compare the perceived outcome?

To assert a "correct" answer to this question is unscientific; just because a participant's judgments of agency deviate from the ground truth defined not by them but by the experimenter, does not mean that their subjective experience was incorrect. In practice, there are many levels of abstraction at which an organism must consider the (lack of) congruence between actions and outcomes to produce effective behavior, and representations at any of these levels may impinge upon the conscious sense of agency. The aims of this dissertation, then, are (1) to elaborate the elements necessary to infer action-outcome pairs in the interest of maintaining a learned sensorimotor behavior, (2) to establish that "low level" sensorimotor processes indeed impinge upon the subjective experience of agency, and (3) to assess the degree to which the sense of agency putatively arising from sensorimotor processes is the same agency one refers to when stating declarative beliefs about their self-agency.

I am, of course, not the first to suggest that the neurocognitive processes underlying sensorimotor control may be involved in generating the sense of agency. It has long been understood that the motor system anticipates the predictable sensory consequences of its

own actions, manifest in inhibitory "corollary discharge" projecting from motor to sensory cortices (Crapse and Sommer, 2008). As such, it is no surprise that this same mechanism was the first proposed to differentiate self-caused from other-caused sensations not just in the early sensory cortical response but also in subjective experience (Feinberg, 1978; Frith, 1987). This early "comparator model" proposal, which had aimed to explain first rank symptoms of schizophrenia, has inspired decades of empirical work suggesting that delusions of control are explained in part by failures to precisely predict the sensory outcomes of actions (Frith, 2012). Ironically, though the comparator model was originally proposed as an explanation for "thought insertion" in schizophrenia, in which a person experiences a thought as coming from another agent, it has failed to do so largely because the distinction between an internal prediction of a thought's content and an actual thought it poorly defined, leaving nothing to compare (Frith, 2012). Modern accounts have addressed this shortcoming by expanding the comparator model into a framework where sensorimotor prediction error is integrated with additional cues prospective (e.g. choice fluency) and retrospective (e.g. outcome valence) cues to produce a final agency judgment (Synofzik et al., 2008; Haggard, 2017). While such a multifactorial framework is appealing in principle and perhaps even likely, in practice researchers may apply the basic comparator model when it works and then liberally append non-motor cues whenever exceptions are encountered. Indeed, though the literature is rife with dissociations between comparator-predicted suppression of the early sensorineural response to a stimulus and perceived agency (Voss et al., 2006; Kühn et al., 2011; Timm et al., 2016), this would-be falsifying dissociation can be circumvented by invoking the idea of a "high-level" comparator that is distinct from the "low level" corollary discharge mechanism documented in the motor control literature. By a different route then, modern iterations of the comparator model – though inspired directly from the motor control literature! – also seem to imply a certain separation between motor control and awareness.

An alternative antidote to the shortcomings of the comparator model could be, rather

than positing successful prediction of the content of sensory feedback as the sole driver of sense of agency, to carefully consider which other mechanisms of motor control may inform judgments of self-causation. In fact, I argue prediction of sensory content is an insufficient basis by which to identify the sensations which arise from one's own movements in the first place. As pointed out by others (Wen and Haggard, 2020), a content-specific action-outcome mapping is unavailable until learned from sensorimotor experience, but one cannot learn an action-outcome mapping without assigning sensory outcomes to actions – leading to a chicken vs. egg style problem. In Chapter 1, drawing on insights from the birdsong model system, I outline the elements of a hypothetical neural circuit that would assign sensory outcomes to specific motor actions, while emphasizing temporal coherence between central (i.e. neural) and peripheral (i.e. biomechanical) events as an organizing principle rather than prediction of content. I present electrophysiological evidence for such a mechanism during human speech production. This finding supports the existence of multiple, likely complementary mechanisms by which action-outcome associations can be identified by the nervous system. Any sensorimotor mechanism that distinguishes sensations caused by action could potentially inform the conscious sense of agency; there is not an *a priori* reason that corollary discharge would have such precedence in awareness.

In Chapter 2, I manipulate the temporal relationship between central motor activity and its (ordinarily) resultant muscle movements to test whether otherwise identical (i.e. same sensory content) movements may be experienced as differing in agency. This is accomplished using a novel Bayesian optimization paradigm, which identifies a latency of electrical muscle stimulation in which participants' neurogenic muscle movements are reliably preempted by an electrically-actuated movement, but they judge roughly half of these externally-caused movements as self-caused. I show that agency judgments can be reliably predicted from the very early sensorineural response to muscle stimulation, measured via electroencephalography (EEG). This result re-implicates "low-level" sensorimotor processing in the experience

of agency over one's body. While Chapter 2's findings firmly links sensorimotor processing to the conscious sense of agency, it is worth noting that our finding pertaining to the EEG response to a muscle movement — prediction of agency judgments from the initial sensorineural response — is not replicated with neural responses to more distal sensory consequences of action, such as a tone following a button press (Voss et al., 2006; Kühn et al., 2011; Timm et al., 2016). This discrepancy begs the question of whether the mechanisms underlying agency are specific to the level to which one is consciously attending. In Chapter 3, we correlate participants' declarative beliefs about the sense of agency, assessed using a validated scale measure, with individual differences in a sensorimotor control detection task. We do, in fact, find such a correlation between participants' psychometric threshold for control detection and their agency beliefs, but we also find that relationship is primarily mediated by the sensitivity of explicit metacognition. That is, it is likely that the sensory evidence of control which is accessible to conscious introspection is preferentially involved in influencing high-declarative beliefs about one's own agency. These findings establish that senses of agency, as defined at multiple levels of behavioral organization, are indeed related constructs. However, a relatively low correlation mediated metacognitive sensitivity suggests that the sense of agency nonetheless invokes distinct cues for judgments made at differing scales of abstraction.

CHAPTER 1

TIMING OF SPEECH IN BRAIN AND GLOTTIS AND THE FEEDBACK DELAY PROBLEM IN MOTOR CONTROL

1.1 Introduction

Speech production requires coordination of over a hundred muscles to achieve desired acoustic output (Simonyan and Horwitz, 2011). Were auditory feedback immediate, acoustic outcomes could be assigned to articulatory movements in real-time, allowing control policies to be learned in a "model-free" manner as in contemporary policy-based reinforcement learning systems (Kakade, 2001). However, muscle contraction, sensory transduction, and intervening multiple stages of neuronal transmission for both all result in intrinsic delays of the arrival of feedback to forebrain structures often associated with learned movements (Wolpert and Ghahramani, 2000; Wolpert et al., 2011). Consequently, vocal learning – and motor learning in general – is thought to require a "forward" model to actively predict current and future peripheral states and, in turn, the resulting sensory feedback from the recent motor output; discrepancies between predicted and actual feedback are used to update the forward model, which can then be leveraged offline to adjust an "inverse" model (i.e. control policy) which proposes movements to achieve desired states (Tourville and Guenther, 2011). This computational account has become a guiding framework for understanding both normal motor behavior and the symptomology of speech motor deficits such as stuttering (Bradshaw et al., 2021). But while forward model updates have been proposed to occur through reinforcement learning mechanisms (Sutton and Barto, 2018), the task of assigning outcomes to actions when feedback is temporally delayed – the hindsight credit assignment problem – remains a standing challenge in theoretical and applied reinforcement learning (Harutyunyan et al., 2019). In particular, when biomechanical dynamics at the body's periphery are nonlinear, the critical events that generate sensory feedback may be temporally decoupled from the

onsets of the muscle movements that caused them; in this case, motor neuron activity would not be separated from sensory feedback by a constant lag, and temporal coherence between the two would not be available as a cue for hindsight credit assignment.

Speech production is one such case. The physics of the vocal organ, the glottis, are such that acoustic output is highly discontinuous with respect to the physiological parameters – subglottal pressure and laryngeal tension – that are directly controlled by the brain (Titze, 1988; Sitt et al., 2008). In particular, in physical models of the glottis – or of the syrinx in songbirds, which has shared dynamics (Amador and Mindlin, 2014) – the position of the vocal folds oscillates given certain pressure/tension combinations but settles at a steady-state position otherwise; when pressure and tension crosses the boundary from a region that yields steady-state dynamics to one with oscillatory dynamics, the oscillations begin abruptly (see Fig. 1.1). These oscillations, called phonation or voicing, are what generate the pressure waves that pass through the upper vocal tract to be shaped into speech. Thus, while onsets of phonation and other discrete acoustic events at the glottis – referred to here as "glottal event onsets" – are directly caused by the motor output of the corticobulbar (laryngeal tension) and corticospinal (respiration/pressure) tracts, their correspondence to actual muscle movements is highly nonlinear.

Some anticipatory mechanism, then, would be required for central neural activity to predict the timing of acoustic/phonatory events in the glottis. However, temporal coherence between such anticipatory representations and auditory feedback would greatly simplify the task of associating movements with their consequences. This proposition would potentially explain the longstanding but perplexing observation that playing a sine tone that is amplitude-modulated to match a human participants' speech envelope with a delay has similarly deleterious effects on speech production as delayed auditory feedback (DAF) (Howell and Archer, 1984); such findings are not well explained by speech production models where forward prediction primarily pertains to the frequency content of sound (Guenther

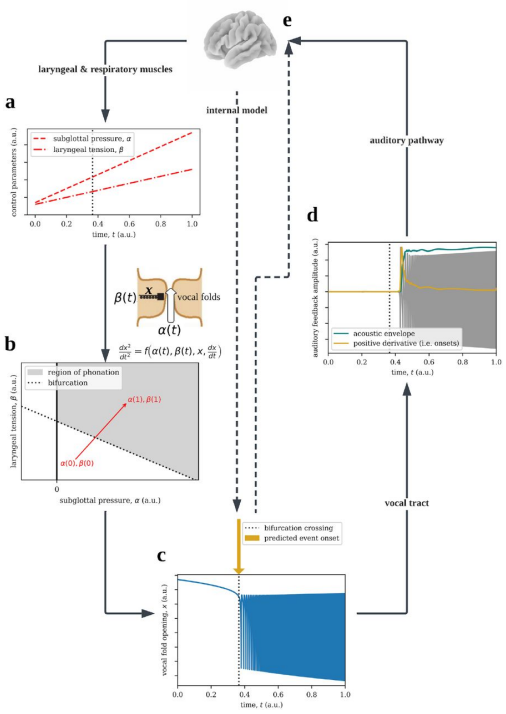

Figure 1.1: The necessity of prediction in the control of voicing. (A) Pictured are arbitrary, continuous trajectories of subglottal pressure and laryngeal tension, which the brain controls through the musculature. (B) In a single-mass model of the glottis, each vocal fold (a.k.a. "vocal chords") are treated abstractly as a mass on a spring, where the stiffness of the spring is set by the tension and an opposing force by the pressure pushing up from the lungs, both interacting with the current position of the folds (e.g. pressure simply passes through the glottis when the folds are open). Per Newton's law, the acceleration will be proportional to the sum of those position-dependent forces, and thus dynamics are governed by a second-order differential equation, which produces oscillatory behavior for some pressure/tension combinations but settles at a steady-state position otherwise (Titze, 1988). (C) When pressure/tension cross the bifurcating border between oscillation-producing and non-oscillatory combinations, phonation begins suddenly, here simulated from the pressure/tension trajectories in panel A using the normal form equations given by Mindlin (2013), which produces sound pressure waves that are filtered through the upper vocal tract to produce speech. To effectively control the onset of voicing, as to produce fluent speech, the brain must anticipate the time of the bifurcation crossing that triggers phonation. (D) The human auditory system is sensitive to the acoustic envelope, but also to the acoustic event onsets that directly result from bifurcation-crossings (Oganian and Chang, 2019), which we call glottal event onsets; however, this auditory feedback reaches the central nervous system at a non-negligible delay, even in normal speaking conditions. (E) If the brain has successfully anticipated the time of glottal events, feedback events will follow those predictions with a fixed lag, allowing sensory feedback to be accurately credited to articulatory events based on temporal coherence. Thus, we hypothesized that glottal event onsets – which we can observe non-invasively via electroglottograph – are reflected in cortex, separately from sensory feedback, during speech production even though they do not directly resemble neuromuscular output.

et al., 2006; Golfinopoulos et al., 2010) . Consistent with this theory, auditory responses in human superior temporal gyrus prominently reflect acoustic event onsets (Oganian and Chang, 2019). However, cortical tracking of corresponding glottal event onsets – which can be measured noninvasively using electroglottography (EGG) – has yet to be demonstrated, which was the aim of the present study.

1.2 Methods

1.2.1 Methods Summary

We recorded subjects' electroencephalography (EEG), audio, and glottis vibrations using electroglottograph (EGG) during speech production. We fit encoding models to test whether glottis event onsets explain EEG variance beyond previously identified features. Further, we exposed subjects to a period of DAF to, in principle, temporally shift their forward model predictions, reflected behaviorally in their threshold for detecting auditory feedback delays (Yamamoto and Kawabata, 2014). Finally, we tested whether a model trained to decode glottal events from brain activity prior to prolonged DAF exposure makes temporally shifted predictions after this threshold shift, reflecting subjects' own updated predictions.

Throughout the experiment, subjects heard their auditory feedback through insert earphones while they spoke into a microphone. Subjects read sentences that are presented on a monitor in front of them during (1) a baseline block with no delay, i.e. normal speaking conditions, and (2) a pretest block in which we imposed a random delay between 0 and 250 ms in their auditory feedback on each trial. To assess whether glottal event times were encoded in the EEG, we trained a multivariate encoding model that predicted the continuous EEG from the speech audio envelope, speech audio onsets, the EGG envelope, and the EGG onsets, and we compared its predictive performance to a model that that excluded EGG audio onsets. Encoding models were trained during the pretest block, where audio and

glottal events were temporally decoupled by design, and tested during the baseline block to assess generalization performance in normal speaking conditions.

Subjects then completed (3) an exposure block with a constant 200 ms delay, and (4) a posttest block that was identical to the pretest block. During the pretest and posttest blocks, subjects were asked to report, on each trial, whether they detected a delay in their auditory feedback; a multilevel logistic model was fit to find the threshold at which they could detect a delay 50% of the time in both blocks, so we could assess perceptual shift as a result of DAF exposure. Then, we trained a decoding model to predict glottal event onsets on the pretest block and tested it on the posttest block (and vice versa) to assess whether the predictions of models that achieve high decoding accuracy are systematically affected by exposure to delayed auditory feedback, consistent with decoding performance being driven by an internal representation that is plastic given sensorimotor experience.

1.2.2 Ethics and Recruitment

We recruited 32 university students (ages 18-22; 15 males, 17 females) over our department's study recruitment system. To obtain the desired level of precision for behavioral measurements, we used the following *a priori* criteria to determine sample size adaptively: subjects were recruited until the 90% HDPI of the Bayesian posterior for pre-to-post shift in delay detection threshold was less than 20 ms (regardless of whether the sign of the shift or whether the HDPI contained zero), which occurred after 29 subjects were collected, after which the additional 3 subjects who had already signed up for the experiment were run and data collection was concluded. Subjects self-reported being right handed, though 3 were considered ambidextrous by the Edinburgh Handedness Inventory (Oldfield, 2013). Subjects were excluded during recruitment if they reported having been previously diagnosed by a clinician with a hearing or speech deficit. Informed consent was obtained prior to the experiment, and procedures were approved by the Social and Behavioral Sciences Institutional Review

Board at the University of Chicago (IRB21-1402).

1.2.3 Experimental Design

After applying EEG and EGG electrodes to the subjects' heads and necks, respectively, they sat in front of a monitor. In each trial of the subsequent task, subjects were asked to read one of the phonetically balanced Harvard Sentences, which was displayed on the monitor (Rothauser, 1969); sentences displayed were on average 9.1 (SD 1.2) syllables long and presented all at once. We controlled the trial-by-trial latency of auditory feedback, recorded through a microphone (Behringer ECM-8000) and played back through insert earphones (Etymotic ER-3C) which blocked out external sound, using a custom Python tool. Since such a system requires digitizing audio and then converting back to analog to play back to the subject, there is an intrinsic minimum hardware/software delay – that is, the actual delay between input and output audio when we specify a delay of zero – which we empirically measured as 4 milliseconds by recording input and output audio simultaneously on our EEG amplifier. For comparison, the intrinsic latency of other auditory feedback manipulation systems used in the literature tends to exceed 15 milliseconds (Kim et al., 2020). The delays recorded in our open dataset (and used in our analysis) have been corrected to account for this minimum delay (by simply adding 4 ms to the value of the "delay" field in the file where delays are recorded).

The task consisted of four blocks: (1) a baseline block in which subjects read 20 sentences with no imposed feedback delay, (2) a pretest block in which subjects read 50 sentences each with a random feedback delay, uniformly distributed between 0 and 0.25 seconds, (3) an exposure block in which subjects read 80 sentences with a constant 200 ms delay, and (4) a posttest block, identical to the pretest block. After each trial in pretest and posttest blocks, subjects were asked to indicate on a keyboard if they detected a delay between their movements and the sound of their voice. Sentence prompts were unique for each subject but

were always phonetically balanced.

1.2.4 Statistical Analysis

From both the delayed speech audio and the EGG, we computed spectrograms between 30 and 5000 Hz using a gammatone filter bank, which approximates the frequency response of the human cochlea and has been noted to outperform other sound frequency decompositions when predicting EEG from the speech envelope (Issa et al., 2024). The envelope of both the audio and EGG was computed by averaging the amplitude across equivalent rectangular bandwidths (i.e. an approximation to the frequency bandwidths in human hearing). As in prior EEG and electrocorticography work, an acoustic event onset time series was computed by taking the positive first derivative within each equivalent rectangular bandwidths and then averaging across them (Oganian and Chang, 2019; Brodbeck et al., 2023). Envelopes (but not onsets) were log-transformed, which better reflects the psychophysical curve for loudness and has been shown to improve prediction of EEG (Brodbeck et al., 2023). Subsequently, all features and the EEG were scaled to be of roughly equal variance using the RobustScaler in the *scikit-learn* package (Pedregosa et al., 2011). Feature scaling was performed within blocks, rather than across the whole dataset, to prevent train-test leakage during cross-validation.

For multivariate EEG encoding models, we estimated Temporal Response Functions as the coefficients of a linear ridge regression that predicts each electrode from values of the time-lagged audio and EGG envelopes from 0.5 seconds before to 0.5 seconds after each EEG time point using the *MNE-Python* package (Gramfort et al., 2014; Crosse et al., 2016). The regularization parameter for ridge regression was chosen using a grid search (between 0.1 and 1e7, log-spaced) to maximize leave-one-trial-out performance on the training set. While it has recently become popular in the EEG literature (Brodbeck et al., 2023), this approach has a longer history in single-cell physiology (Theunissen et al., 2001) and is also frequently

used, though referred to as "voxelwise-encoding," in the fMRI literature (Huth et al., 2016; Dupré la Tour et al., 2022). Encoding models were trained on the pretest block, where the randomly varied latency of the auditory feedback allowed the model to differentiate between the EGG and audio waveforms, which would otherwise be roughly synchronous, and then tested (i.e. cross-validated) under normal speaking conditions in the baseline block, where we computed the correlation between the predicted and actual values at each electrode to measure the encoding models' generalization performance. We fit both an encoding model that includes all four audio and EGG features and one excluding EGG onsets to test whether EGG event onsets explain variation in brain activity beyond that already explained by the audio envelope, audio onsets, and the EGG envelope. Cross-validated correlations of both models' predictions with each EEG electrode were statistically compared to chance using a permutation test where null permutations are generated by shuffling the predicted time-series across trials, and to each other using a paired permutation test, both corrected for multiple comparisons using threshold-free cluster enhancement (with clustering across neighboring electrodes) to control the familywise error rate (Smith and Nichols, 2009). Univariate (i.e. single feature) encoding models as in Fig. 1.2 were constructed by directly inverting our decoding models as described by Haufe et al. (2014), such that they correspond exactly to the same patterns selected by our decoders (Haufe et al., 2014). We compare the electrode-by-time weights of the univariate encoding models to zero using the t-max procedure with 10,000 permutations to control for multiple comparisons, and we show the encoding model weights averaged across subjects in Fig. 1.2 (Nichols and Holmes, 2002). We also trained a separate encoding model for glottal event onsets on the posttest block, and we compared the weights of the pretest and posttest encoding models using a spatiotemporal cluster-based permutation test (Maris and Oostenveld, 2007).

Decoding models were fit using the exact same procedure (same code) as the encoding models but in the reverse direction, predicting the values of each audio and EGG features

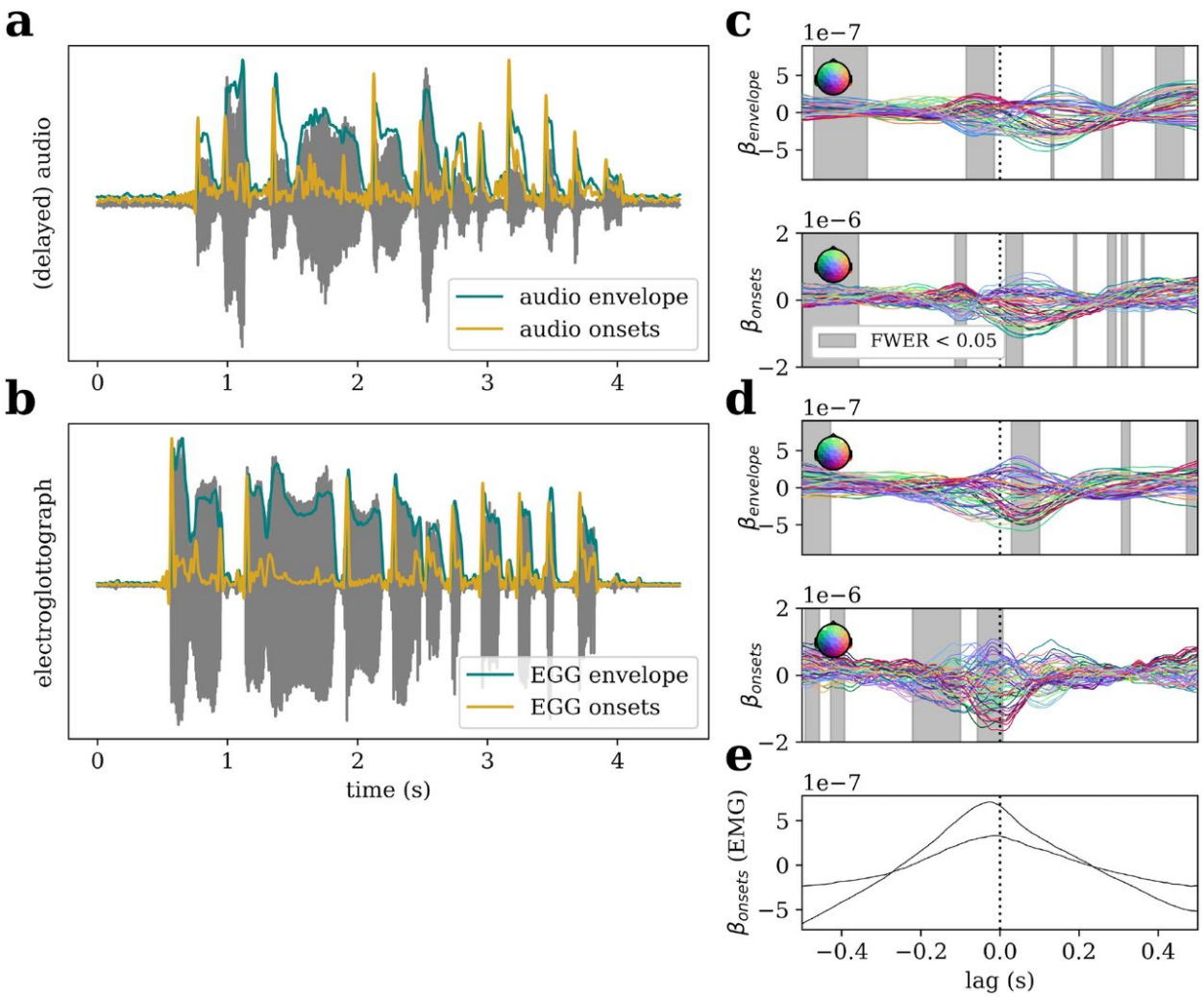

Figure 1.2: EEG temporal response functions to features of glottal waveform and auditory feedback waveform during speech production. From (A) the speech audio, played back to the subject at a delay subsequent to the baseline block, and from (B) the glottal waveform, we compute an amplitude envelope that approximates the frequency response of the cochlea and an "onset" time series that reflects the positive first derivative of the amplitude across all frequency bands – thus the onset of spectral events. Temporal response functions (TRFs), the EEG responses to these features of the (C) auditory feedback or (D) glottal waveform, are shown; a large coefficient in the TRF at, for example, lag 0.1 s means that a change in the value of that feature manifests in the EEG 0.1 seconds after that change. Shaded intervals cover time lags at which model weights for at least one electrode deviate from zero, after correcting for multiple comparisons. Model weights for encoding of glottal onsets are significant throughout an interval up to and including zero lag, consistent with neural activity anticipating glottal events, rather than merely responding to sensory feedback. (E) For comparison, a TRF to glottal onsets fit to the jaw EMG, rather than scalp EEG, channels. The predictions of the corresponding EMG-based decoding model are used as a control variable in subsequent EEG decoding analyses.

from the -0.5 to 0.5 second time-lagged EEG electrodes. (That is, decoding models are like fitting an encoding model to the speech features, with the EEG electrodes as predictors.) While only the decoding of the EGG onsets is of interest in the present study, we report the cross-validated decoding performance – again, trained on pretest and tested on baseline – for all features, with p-values comparing each to chance using a permutation test where null permutations are generated by shuffling decoded time-series across trials.

Since EEG artifact correction methods can only be expected to attenuate, not completely remove, non-neural artifacts from the EEG signal, we performed a mediation analysis to assess statistically the extent to which EMG contamination may have explained successful decoding of glottal event onsets. To do this, we trained another decoding model to predict the glottal onsets directly from the two EMG channels on the pretest block, and generated predicted event onsets for the baseline block (exactly as we had with the EEG channels). For each subject, we then fit a mediation model for the linear relationship between the EEG-predicted onsets and the actual onsets, with the EMG-predicted onsets as a mediator; this gave us a per-subject "pure" estimate of the direct effect of the EEG in predicting onsets (left over when controlling for the confounding influence of EMG), and a separate estimate of the indirect effect which could be explained by EMG contamination. We compared the group-mean direct effect to chance with a permutation test (again shuffling the EEG-predicted time series across trials). We also performed a permutation test separately for each subject and fit a p-curve mixture model to the resultant p-values, which allowed us to use model comparison to assess relative evidence for population prevalence models in which there is a nonzero direct effect in none of, all of, or merely some subjects in the population (Veillette and Nusbaum, 2025); specifically, Bayesian model comparison is performed using the expected log pointwise predictive density (ELPD) with Pareto-smoothed importance sampling leave-one-out-cross-validation (Vehtari et al., 2017). We report bootstrap 95% confidence intervals for the proportion of the total effect that is accounted for by the direct effect.

To manipulate the content of subjects' forward models, we exposed subjects to 0.2 second delayed auditory feedback throughout the exposure block, after which the same procedure as the pretest block was repeated in posttest. We estimated the shift in subjects' threshold for detecting auditory-motor asynchronies, where "threshold" is defined as where the probability of detection is 50%, using a Bayesian multilevel logistic regression estimated with the *PyMC* package (Patil et al., 2010). Under the assumption that subjects detect a delay when their auditory feedback is sufficiently different from the feedback they predict, this threshold is a behavioral index of the content of their forward model and, as a corollary, a pretest to posttest change in threshold indicates a change in their forward model. Since delaying auditory feedback impairs speech production, usually resulting in slowing of speech or other disfluencies, we also fit a multilevel Poisson regression for speech rate as a function of delay and any change in that function from the pretest to posttest block. The speech rate on each trial was measured as the number of syllables in the target sentence divided by the duration of voicing, which was quantified in *Praat*.

If the EEG signal that reflects EGG event onsets is indeed an index of predictive cortical tracking of glottal events then one might expect that signal should change from pretest to posttest as do subjects' detection thresholds. To test this hypothesis, we take those models which have been trained to predict EGG event onsets in the pretest block, and we generate its predicted onset time series for the posttest block; we then compute the time lag at which the cross-correlation between the predicted and actual event onsets is maximal as an estimate of how much predictions have shifted in time between pretest and posttest. Additionally, we train a decoding model on the posttest block and compute the same cross-correlation lag testing on the pretest block.

These measurements allow us to assess the possibility that the signal driving decoding is (a) a response to the glottal waveform via another sensory modality, e.g. somatosensory, or through bone conducted auditory feedback, or it reflects (b) a movement/muscle artifact

or contamination from the EGG recording itself. Should either of these possibilities account for above-chance decoding performance, then there should be no trend in the central tendency of pre-to-posttest prediction lags or post-to-pretest lags with increasing decoder performance; at-chance models will always average zero lag, since any peak in the cross-correlation function would be spurious, and higher performing models will tend toward zero lag as signal-to-noise for decoding from the confound signal, which should not be affected by perceptual recalibration following DAF, improves. Conversely, if the signal that drives decoding performance is related to subjects' forward model, then the predictions of decoders with higher signal-to-noise should be more affected (since low performing models will still be randomly distributed about zero lag, as their predictions are essentially random). Thus, the alternative hypothesis predicts a monotonic trend in pre-to-posttest lags and an opposite trend in post-to-pretest lags as baseline decoding performance improves. Thus, we test the null hypothesis using a Spearman rank correlation between the cross-validated correlation (trained on pretest, tested on baseline) for each subject and their pre-to-posttest shift in decoded onsets. We additionally report rank correlations between the lags and the direct (EEG) and indirect (EEG) effect estimates.

It is not necessarily the case that a shift in detection threshold corresponds directly to a same-direction shift in the expected time of glottal events as predicted by subjects' forward models; psychophysical evidence indicates that humans also maintain a prior on the latency between action and sensory outcomes, and whether the expected movement time or expected action-outcome latency is updated following a prediction error would depend on the relative prior precision of the two estimates (Legaspi and Toyoizumi, 2019). However, the former explanation does make the distinct prediction that that pretest-trained decoder predictions will be delayed in the posttest block, and posttest-trained decoder predictions will be early in the pretest block. To test this additional prediction, we aimed to quantify the average temporal shift in decoded glottal event onsets from before to after adaption

while controlling for the confounding influence of EMG contamination, which is indexed by the indirect effect estimate from the mediation analysis described above. To do so, we fit a robust linear regression of the form lag $= \alpha_0 + \alpha_1$(indirect effect), such that the intercept α_0 is interpreted as the expected value of the lag when the indirect effect is equal to zero. We report the regression p-value and confidence interval for this intercept term, fit to (a) just the pre-to-post lags and (b) the average lag for each subject, where post-to-pre lags have been flipped so their sign has the same interpretation as the pre-to-post lags.

1.2.5 EEG Acquisition and Preprocessing

EEG was acquired at 10,000 Hz using an actiCHamp amplifier (Brain Products, GmbH) with 60 active EEG electrodes, two electrooculogram (EOG) electrodes below the eyes, and two electromyogram (EMG) channels over the jaw muscles in front of the ears, referenced to their ipsilateral mastoid. The precise positions of each EEG electrode were measured and recorded using a CapTrak (Brain Products, GmbH).

Since subjects were speaking during the study, much care was taken to attenuate contamination of the EEG signal from facial muscles. EEG was first processed using the PREP pipeline, which performs bad channel identification/interpolation, re-referencing, and line-noise removal in a standardized, robust manner (Bigdely-Shamlo et al., 2015; Appelhoff et al., 2022). Subsequently, we bandpass-filtered the EEG between 1-70 Hz and then decomposed the data into 60 independent components (ICs) and removed components that were significantly correlated with the EOG channels. We also, importantly, removed ICs that showed evidence of facial EMG contamination using the spectral slope, peripheral (i.e. far from vertex) power, and focal point (i.e. low spatial smoothness) criteria previously validated by Dharmaprani and colleagues using data from paralyzed patients (Dharmaprani et al., 2016). After removal of artifactual ICs, we then epoched the data and dropped trials that still showed evidence of strong non-neural signal contamination (i.e. peak-to-peak

voltage exceeding 150 microvolts). To address any subtler movement artifacts, we then computed the current source density transformation of the EEG data using a spherical spline surface Laplacian. This transformation, effectively a spatial high pass filter, attenuates the effects of volume conduction across the scalp such that each electrode predominantly reflect nearby physiological sources (Kayser and Tenke, 2015). Previous work, again using data from paralyzed patients, has shown that the surface Laplacian effectively removes EMG contamination from central and peri-central electrodes (Fitzgibbon et al., 2013). Finally, the EEG was down-sampled to 140 Hz (i.e. twice the lowpass cutoff) to reduce the computational burden of our analysis.

All EEG preprocessing was automated to ensure reproducibility – using *pyprep*'s implementation of the PREP pipeline and the *MNE-Python* implementation of ICA and routines to select artificial ICA components –but data quality was manually verified after the PREP pipeline and again after artifact correction. In addition to the preprocessed data itself, detailed preprocessing reports that were generated at these points are available for each individual subject as part of our open dataset. These reports contain power spectra of the EEG signal after PREP, a list of electrodes that were interpolated (and the criteria by which they were flagged during PREP), visualized scalp topographies of all removed ICs, and remaining trial counts for each condition after rejection criteria are applied.

1.2.6 EGG and Audio Acquisition and Synchronization

EGG was acquired using an EG2-PCX2 electroglottograph amplifier (Glottal Enterprises, Inc.) and audio using a Behringer ECM-8000 microphone with a roughly flat frequency response. These signals were sent from the EGG amplifier to a StimTrak (Brain Products, GmbH) via an audio passthrough cable, and were then digitized by our EEG amplifier so that all signals were recorded on the same clock. Like most EGG systems, the EG2-PCX2 has a built-in highpass filter, which introduces phase-specific delays into the recorded

signal. To reverse this phase delay, we applied a digital equivalent of this 20 Hz highpass filter backwards after acquisition. (This sequential forward-backward filtering is the same operation that "zero phase shift" filters, like *scipy*'s and MATLAB's *filtfilt* functions, apply.)

1.2.7 Data and Code Availability

All raw data and preprocessed derivatives, included fitted models, are available on OpenNeuro (https://doi.org/10.18112/openneuro.ds005403.v1.0.1) and are organized according to the Brain Imaging Data Structure conventions for EEG (Pernet et al., 2019). Versioned analysis code is archived on Zenodo (https://zenodo.org/doi/10.5281/zenodo.13238912), and task code is available on GitHub (https://github.com/apex-lab/daf-experiment).

1.3 Results

1.3.1 Glottal waveform events reflected in the EEG during speech production

The encoding model that incorporates all features – speech audio envelope, audio onsets, EGG envelope, and EGG onsets – performed robustly above chance (TFCE statistic = 11.34, p = 0.0001), as did the model that excludes EGG onsets (TFCE = 10.12, p = 0.0001). However, including the EGG onsets in the model did improve out-of-sample prediction of the EEG signal (TFCE = 2.38, p = 0.0085), indicating that the time of events in the EGG explains additional variance in cortical neural activity that is not explained by these other features.

The EEG temporal response function (TRF) to EGG onsets, visualized in Fig. 1.2, shows significant model weights leading up to and including the actual time of the glottal event. This is too early to be caused by sensory feedback (which, by definition, occurs after the event) and aligns with the theoretically predicted timescale of a forward model's predictive representations, evidence of which is observed both in the motivating songbird studies and,

though outside of the auditory system, in primates (Mulliken et al., 2008; Amador et al., 2013; Dima et al., 2018). However, a TRF time-locked to movement could also be predicted by either a movement artifact or perhaps by a motor command precipitating the movement itself; our subsequent decoding analyses were designed to assess the possibility of such confounds.

1.3.2 Glottal event onsets decoded from the EEG

All four features, the envelope and onsets for audio and EGG signals, could be decoded from the EEG with above-chance cross-validation performance, with all $p = 0.0001$, the minimum possible p-value for our permutation test. The out-of-sample correlation [95% CI] between decoded and actual feature time series was $r = 0.233$ [0.199, 0.268] for the audio envelope, $r = 0.101$ [0.086, 0.117] for audio onsets, $r = 0.349$ [0.312, 0.382] for EGG envelope, and $r = 0.136$ [0.120, 0.152] for EGG onsets. The full distribution of subject-level decoding performances is shown in Fig. 1.3a, next to the encoding model performances in Fig. 1.3b-c. Moreover, when we control for EMG contamination in a mediation analysis, we estimate that 74.5% [65.8, 80.8] of the total linear relationship between EEG-predicted glottal onsets and actual onsets is accounted for by a direct effect (i.e. not by EMG), and our permutation test found this direct effect was significant at the group level ($p = 0.0001$). We also performed a permutation test for each subject individually to estimate the population prevalence of the effect with a p-curve mixture model (Veillette and Nusbaum, 2025); Bayesian model comparison found that data were best explained by a model in which all subjects (rather than just some) showed a direct effect (ELPD = 153.2), assigning a weight of 0.95 to that model. (This is interpreted loosely as a 0.95 posterior probability that this model would best predict the results of an additional subject.) In other words, formal prevalence inference suggests this effect may be nearly universal in the population from which we sampled (Veillette and Nusbaum, 2025).

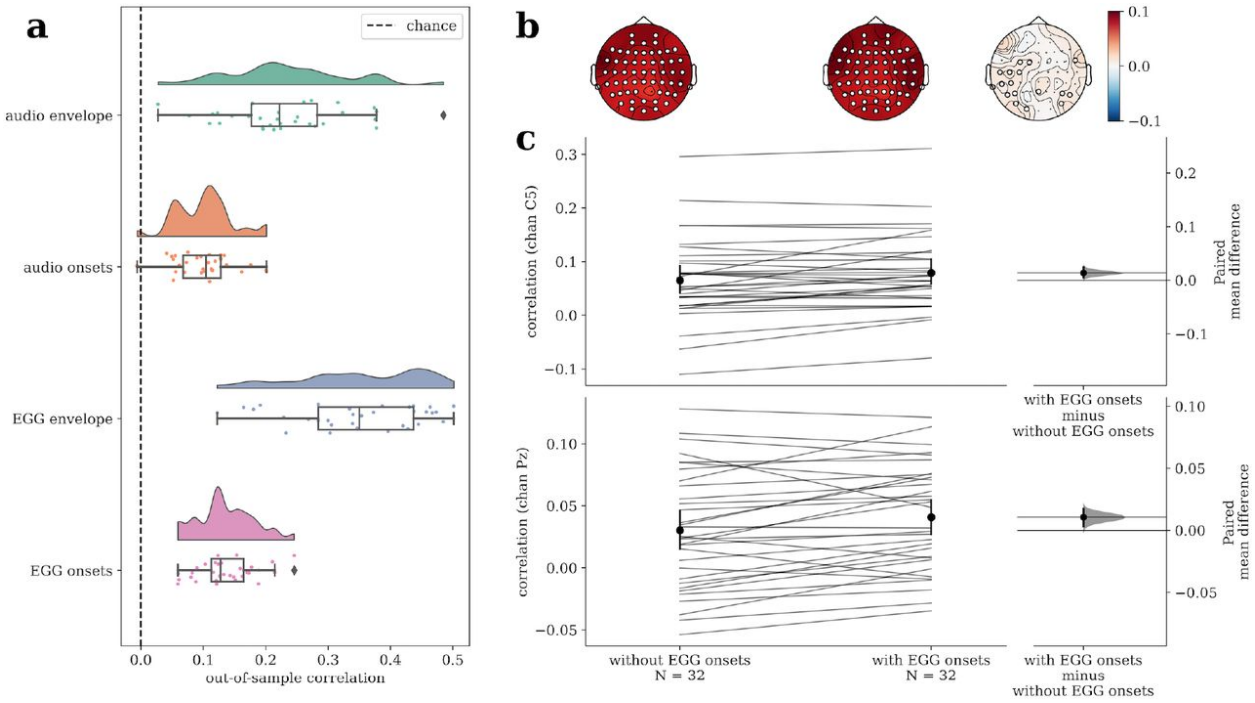

Figure 1.3: Out-of-sample predictive performance of decoding and encoding models.(A) Per-subject cross-validated Pearson correlations between actual time series of the audio/EGG features and those decoded from the EEG are shown as individual points, with overlaid boxplots showing quartiles of the observed correlations and whiskers extend 150% the interquartile range past the upper and lower quartiles to show the realistic extent of the distribution. Diamonds represent subjects falling outside of this range, and kernel density estimates are shown above. (B) The cross-validated Pearson correlation between actual EEG and that predicted from the audio/EGG features, shown for all electrodes. The model on the left excludes the EGG onsets as a predictor, the middle includes EGG onsets, and the difference in performance is shown on the right. White circles mark electrodes that are significant after multiple comparisons correction. (C) The same encoding model performances are shown for a central (C5) and a parietal (Pz) electrode; paired correlations for each subject, their bootstrap means with 95% confidence intervals, and the bootstrap distribution of the mean differences are shown. All decoding and encoding models visualized here were trained in the pretest block (i.e. before exposure) and tested in the baseline block – in other words, in normal speaking conditions.

1.3.3 Temporal recalibration of perception of auditory-motor synchrony after prolonged exposure to delayed auditory feedback

The group-level parameters [95% highest posterior density interval (HPDI)] of the logistic multilevel model (time units in milliseconds) used to measure delay detection threshold were β intercept= -2.19 [-2.64, -1.70], β_{delay} = 36.6 [31.16, 42.25], β_{adaption} = -0.97 [-1.52, -0.414], $\beta_{\text{delay}\times\text{adaption}}$ = 2.61 [-3.18, 8.47]. The detection threshold, or the delay at which the probability of detection is 50%, shifted from mean 59.8 ms pre-exposure to 80.8 ms post-exposure; in other words, subjects' detection thresholds shifted by about 20.9 ms [10.7, 31.6] after the exposure block. Trial-level behavioral data (Fig. 1.4, top), group-level logistic curves, and both group- and subject-level estimates of threshold shift (Fig. 1.4, bottom) are visualized in Fig. 1.4. Recalibration of auditory-motor asynchrony detection thresholds indicates that subjects' internal predictions about the time of articulatory events and/or the latency of their sensory consequences were altered by the DAF manipulation.

In contrast, there was not strong evidence of motor adaption as a consequence of prolonged DAF exposure. A Poisson regression found that speech rates indeed slowed, as expected, as a function of the trial-by-trial delay ($\beta_{\text{intercept}}$= 1.40 [1.35, 1.45], β_{delay} = -0.87 [-1.15, -0.61]). However, the credible intervals of adaption-related changes both contained zero (β_{adaption} = 0.04 [-0.02, 0.010], $\beta_{\text{delay}\times\text{adaption}}$ = -0.03 [-0.38, 0.29]). Taking the upper edge of the 95% credible interval as an upper bound, this suggests any main effect of exposure is smaller than a 0.37 syllables/second difference, if such an effect exists at all.

1.3.4 Decoded glottal event times affected by perceptual recalibration

When we trained decoding models to predict EGG onsets from brain activity on the pretest block, we found that models which performed better in normal speaking conditions (i.e. baseline block) predicted later onset times relative to the actual EGG events when tested during posttest (Spearman's rank correlation = 0.410, p = 0.0198). The temporal shift was

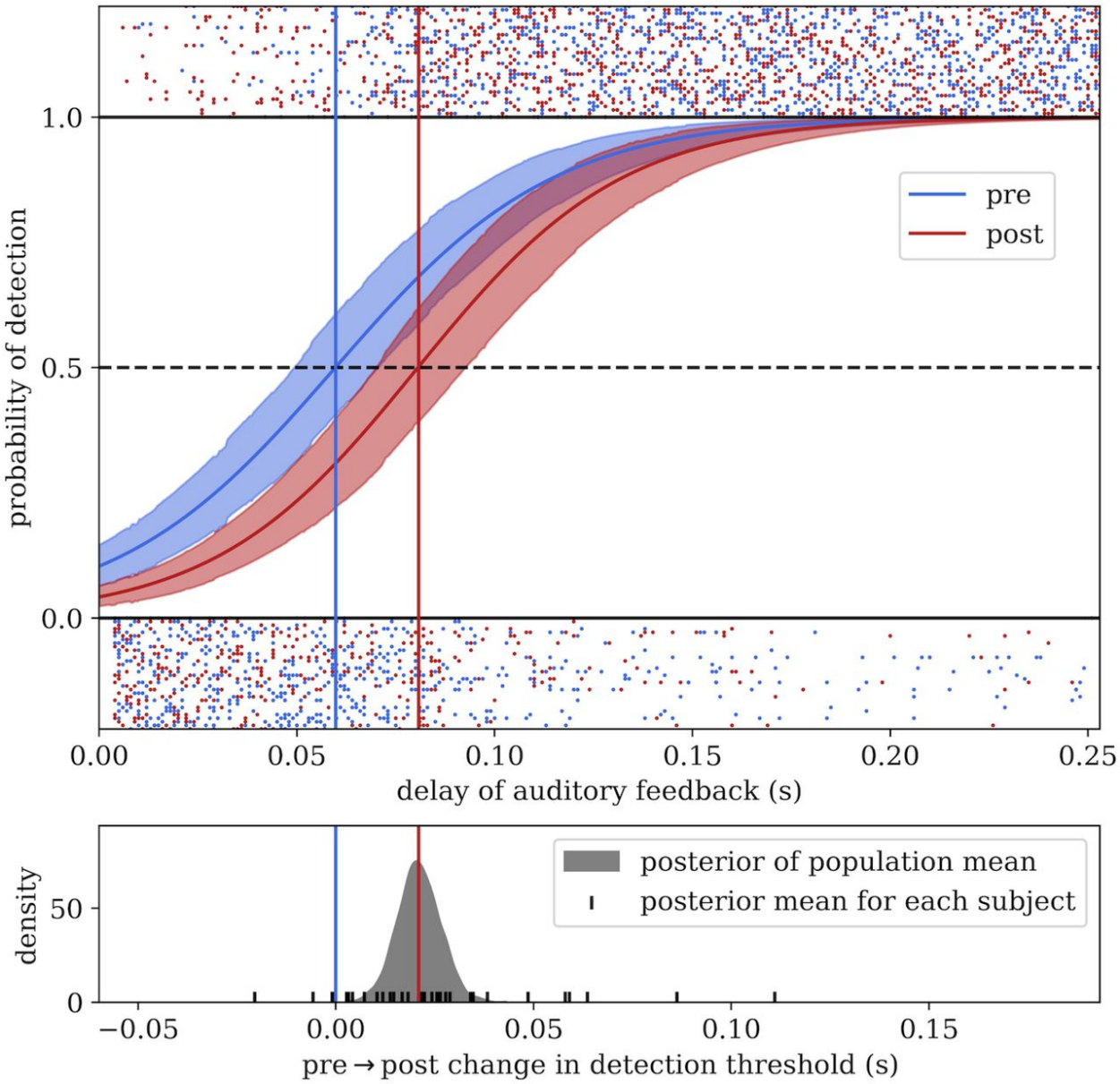

Figure 1.4: Behavioral shift in delay detection threshold following exposure to delayed auditory feedback. (top) The group-level logistic regression functions describing the probability of detecting an auditory-motor asynchrony in the pretest and posttest blocks, which flank an exposure block where subjects are exposed to a constant auditory feedback delay. Trial-by-trial responses are shown above and below the logistic curves, with rows corresponding to subjects. (bottom) The posterior distribution of the population/group average change in the delay detection threshold, with point estimates shown for each individual subject.

also positively correlation with the EEG direct effect (rank correlation = 0.396, p = 0.0251) but negatively correlated with EMG contamination (rank correlation = -0.446, p = 0.0105). In contrast, models trained on models trained on posttest (which resulted in different model weights; cluster based permutation test: p = 0.0039, see Fig. 1.5b) and tested at pretest showed the exactly the opposite pattern, where lags varied negatively with model score (rank correlation = -0.350, p = 0.0497) and with the direct EEG effect (rank correlation = -0.413, p = 0.0189), though not significantly with the indirect effect (rank correlation = 0.192, p = 0.293). The fact that decoder performance is monotonically related to how much a decoder is affected by exposure to an auditory-motor delay rules out the possibility that decoding performance is explained by a movement artifact or non-plastic sensory response, which would be invariant to temporal recalibration. (We do note that, despite this group trend, lags of zero are still common in the sample, so it is certainly possible that such confounds still contribute to decoding performance as also suggested by our mediation analysis – they just cannot fully explain it.)

To estimate the mean temporal lag in predictions due to DAF exposure, we performed a robust regression while controlling for the confounding influence of EMG contamination (see Fig. 1.5d). Using only the pre-to-posttest lags, we estimated that shift as roughly 3.6 milliseconds (p = 0.0359, 95% CI: [0.2, 7.0]). Incorporating both pre-to-posttest and (sign-flipped) post-to-pretests lags as repeated measures, we estimated a slightly smaller but still significant shift of roughly 2.9 milliseconds (p = 0.0483, 95% CI: [0.02, 5.8]). In other words, we do find evidence of a slight positive shift on average, consistent with an internal prediction of event time updating after exposure to delayed sensory feedback. However, we note that the estimated shift is an order of magnitude smaller than the behavioral shifts in the delay-detection threshold, and thus unlikely to fully explain perceptual adaption.

As the decoded signal can decouple in time from movements, it is neither well explained by a motor command as traditionally conceptualized, which would lead the movement with

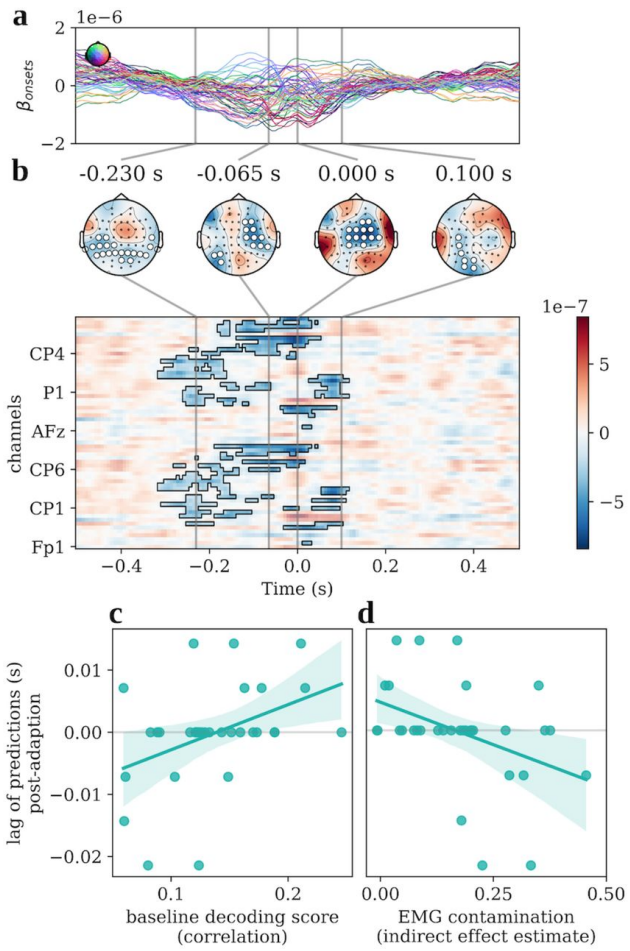

Figure 1.5: Exposure to delayed auditory feedback alters the neural activity tracking glottal event times. (A) The weights of EEG encoding model for glottal event onsets trained during the posttest block. (B) Difference wave for posttest encoding model vs. the encoding model trained during pretest and seen in in Fig. 2d, with spatiotemporal cluster that exceeded change magnitude in cluster-based permutation test highlighted and electrodes marked on scalp. (C) The lag between actual event times and event times predicted by decoder trained in the pretest block and tested in the posttest block, shown as a function of baseline decoding performance and (D) as a function of estimated EMG contamination. High performing models and models that show minimal EMG contamination tend toward positive shifts, suggesting EEG-based predictions are delayed relative to events following prolonged exposure to delayed auditory feedback.

a constant delay, nor by a movement artifact, which would be inherently synchronous to the movement. Rather, high decoding performance appears to be driven by a signal that is ordinarily temporally coherent with glottal articulatory movements but can decouple following exposure to an auditory-motor mismatch. These operating characteristics are indicative of an internal prediction, which is malleable to sensorimotor experience.

1.4 Discussion

The present study presents evidence that neural activity in human cortex tracks the timing of events in the glottal waveform during speech production in roughly temporally coordinate fashion. Acoustic events in the glottal waveform have a lawful temporal correspondence to the subglottal pressure and laryngeal tension gestures that are used to control the vibration of the glottis (Titze, 1988), as is true for the equivalent structures in avian vocal learners (Perl et al., 2011; Mindlin, 2013; Boari et al., 2015). Moreover, the glottal waveform serves as the acoustic source that, once filtered through the upper vocal tract, results in the speech waveform (Titze, 1988). Thus, under normal speaking conditions, event onsets in the glottal waveform have a one-to-one correspondence with subsequent acoustic events in the auditory stream, and such onset events are known to be prominently reflected in the auditory cortical representation of speech (Oganian and Chang, 2019). In other words, cortical tracking of glottal gestural/acoustic event times could contribute to the circuit resolving the temporal credit assignment problem – based on temporal coherence between event times – in the control of voicing (Harutyunyan et al., 2019).

Work in birdsong notably mirrors but also helps to elaborate these observations. In the HVC (premotor/association cortex analog), certain neurons produce spike bursts and interneurons are suppressed within milliseconds of transitions between the syringeal pressure and tension gestures used to control birds' analog to the human glottis within their syrinx (Amador et al., 2013). As bursts occur too late to have triggered movements but too early

to represent sensory feedback, this synchronous activity has been interpreted as a state prediction. More recent work has shown that some primary motor cortex analog projecting HVC neurons also burst at syllable onsets or offsets and at prominent transitions within syllables, although the timing of these burst relative to syringeal gestures has yet to be formally evaluated (Moll et al., 2023). Some models of birdsong production proposed that this sort anticipated synchronization of motor neuron activity with peripheral events emerges organically as a characteristic of certain coupled dynamical systems, which yield zero lag relationships between hierarchically organized structures (Alonso et al., 2015; Dima et al., 2018). The patterns of respiratory activity in singing canaries are also predicted by such models (Alonso et al., 2015; Dima et al., 2018). Notably, pressure-tension gesture transitions are directly reflected in the derivative of the song envelope (Boari et al., 2015), and some Field L (auditory cortex analog) neurons selectively respond to the same derivative information in auditory stimuli (Sharpee et al., 2011). As some HVC neurons are prone to repeat bursts after fixed intervals following their initial recruitment during song, rebound bursts may occur near the time of sensory feedback (Daou and Margoliash, 2020; Fetterman and Margoliash, 2023). These observations provide elements for simultaneous neural representation of gestures and their acoustic outcomes, which could be linked by well-documented coincidence detection mechanisms (Joris et al., 1998; Matell and Meck, 2004).

It is worth elaborating that such a mechanism, emphasizing event onsets and transitions rather than event content as described, complement the content-specific forward model predictions posited by extant models of motor control (Tourville and Guenther, 2011). Frequency-specific suppression of sounds resulting from movement is well established in mammalian auditory cortex (Schneider et al., 2018; Schneider and Mooney, 2018), but to learn such content-specific mappings in the first place, we posit an organism must grapple with the hindsight credit assignment problem. Representing articulatory movements in a manner that is temporally coherent with resultant peripheral events ensures sensory feedback oc-

curs at a fixed lag relative to motor neuron activity, which affords learning content-specific action-outcome mappings via simple associative mechanisms. In speech production, this would predict – as we present evidence for here – a temporal correspondence between some neural activity and the onsets of discrete acoustic events that occur when continuous changes in subglottal pressure and laryngeal tension act on the nonlinear physical dynamics of the voice's sound source, the glottis (Sitt et al., 2008; Boari et al., 2015). In this view, then, hindsight credit assignment when processing auditory feedback is emergently facilitated by the manner in which the central nervous system coordinates with the vocal organ, circumventing the need for specialized CNS mechanisms to temporally realign actions-outcome pairs post hoc.

CHAPTER 2

TEMPORAL DYNAMICS OF BRAIN ACTIVITY PREDICTING SENSE OF AGENCY OVER MUSCLE MOVEMENTS

2.1 Introduction

Voluntary movements are usually accompanied by an experience of "I did that." This feeling is the sense of agency (SoA), which is considered a basic building block of conscious selfhood (Gallagher, 2000; Haggard, 2008). Pathologies affecting SoA, including schizophrenia (Frith, 2012), alien hand syndrome (Panikkath et al., 2014), perceived (non-)control of a phantom limb (Ramachandran and Hirstein, 1998), automatic "utilization behavior" (Lhermitte et al., 1986), and learned paralysis (Wolf et al., 1989), are often characterized by anomalies in the experience of control over the body itself (i.e., the musculature) rather than external action outcomes per se.

SoA research in healthy populations, however, has focused primarily on external consequences of action (Haggard, 2008, 2017). While some studies have manipulated bodily agency by delaying visual feedback from movements, such manipulations leave intact the somatic sensation of muscle movement, over which the subject might still feel agency in the absence of SoA over the decoupled visual stimulus (Tsakiris et al., 2010; Abdulkarim et al., 2023). Others have noted the lack of experimental paradigms addressing "narrow" SoA over muscles as opposed to "broad" SoA encompassing action outcomes (Christensen and Grünbaum, 2018). The field often assumes conclusions drawn from paradigms investigating SoA over a tone following a button press will generalize to other classes of agency judgments. As such, the literature tends to treat SoA as a homogeneous phenomenon always accompanied by the same neural correlates. However, an alternative hypothesis is that the neural correlates of SoA may vary as a function of modality (e.g., proprioceptive vs auditory) or the level of abstraction for a given judgment (Charalampaki et al., 2022). Indeed, it has been

argued current models may not generalize to SoA over the musculature (Christensen and Grünbaum, 2018) or over thoughts (Frith, 2012). This discrepancy bears on a fundamental question of whether mechanisms that give rise to the experience of an agentic self are common across scales of biological and social-behavioral organization — and if not, how and why do we assign agency to the same unified self at these different scales (Veillette et al., 2024)?

One reason for the shortage of paradigms assessing SoA over movements is that control over one's own muscles is normally unambiguous. Indeed, previous attempts to elicit SoA for experimenter-evoked (e.g., by transcranial magnetic stimulation) movements have not succeeded (Haggard and Clark, 2003; Christensen and Grünbaum, 2018), leading to the conclusion that "involuntary movements are never accompanied by a sense of agency" (Haggard, 2017). However, since cognitive scientists favor indirect SoA measures, such as intentional binding (perceived delay) between actions and outcomes, these findings primarily reflect SoA over outcomes rather than SoA over the muscles (Haggard et al., 2002). Meanwhile, human–computer interaction researchers have begun investigating SoA in interfaces that use electrical-muscle stimulation (EMS) to drive users' muscles. They find, in contrast, that subjects report EMS-caused movements as self-caused so long as stimulation temporally aligns with users' endogenous intention to move (Kasahara et al., 2019, 2021; Tajima et al., 2022). Cognitive neuroscientists have yet to embrace these findings, partly because self-reports may result from response biases (Dewey and Knoblich, 2014), lacking convergent validation from neural measurements.

Thus, in the present work, we "preempt" subjects' endogenous movements with EMS during a cue-response reaction time task, using manipulating stimulation timing to control the proportion of EMS-caused movements perceived as self-caused. Using time-resolved decoding of subjects' trial-by-trial EEG, we show that cortical activity predicts agency judgments about resulting muscle movements as early as 83 ms following stimulation, showing

that subjects' self-report has a basis in early, preconscious sensorimotor processing, not just a response bias. Notably, this result differs from those obtained using typical button-tone paradigms, where early evoked responses (to tones) have repeatedly failed to predict subjective agency judgments (Kühn et al., 2011; Timm et al., 2016). Finally, an exploratory analysis shows that fractal measures also predict SoA, suggesting that complexity of sensory processing may differ for sensations perceived as movement feedback.

2.2 Materials and Methods

2.2.1 Methods summary

The goal of our experimental design was to evoke movements using EMS in which ∼50% of such movements were perceived as self-caused (agency) and 50% as EMS-caused (nonagency), although all such movements are indeed EMS-caused. Previous work has shown that EMS-caused movements are perceived as self-caused if they modestly preempt subjects' natural movements in a reaction time task (Kasahara et al., 2019; Tajima et al., 2022), and varying the timing of stimulation has been used to manipulate agency (Kasahara et al., 2021). However, this manipulation results in the stimulation latencies being systematically different between agency and nonagency trials, presenting a clear confound for neural analysis. To this end, we designed a procedure in which stimulation timing is tuned on a per-subject basis to a latency at which subjects report (without further manipulation of stimulation latency) that movements were self-caused on approximately half of trials (see Subsection 2.2.3). This results in maximally similar distributions of stimulation latency across agency and nonagency trials.

In our analysis, we aim to identify patterns in the scalp EEG response to muscle stimulation that robustly predict whether resulting muscle movements are perceived as self-caused or perceived as EMS-caused on a trial-by-trial basis, with a particular focus on the temporal

characteristics of those patterns. First, we train a linear classifier at each time (relative to stimulation onset) throughout the epoch, and test its generalization performance across subjects and across time (see Subsection 2.2.4). An advantage of this approach is that it gives us information not just about when patterns that predict agency emerge, but how long those pattern remains present and continually predictive (King and Dehaene, 2014). This allows us to differentiate, for instance, patterns of neural activity that appear only transiently from those that are sustained over time. In addition, we assess whether complexity measures of the EEG response — the fractal dimension, and index of signal complexity, and the Hurst exponent, and index of long-range temporal dependency — predict trial-by-trial SoA. These complexity features allow us to uncover some of the qualitative characteristics of neural dynamics (e.g., sensitivity to perturbation, scale-freeness, or self-similarity) in the presence and absence of SoA, although it should be noted since this latter analysis was exploratory, its evidential value should be weighted accordingly.

2.2.2 Participants and ethics statement

Twenty-five University of Chicago undergraduate students (6 male, 19 female, ages 19-24 years) participated in the study; however, 2 subjects were subsequently excluded for non-compliance with task instructions (i.e., one admitted to letting the electrical stimulator perform the task without attempting a volitional button press, and another pressed the button continually to speed through the task instead of when cued). Participants were recruited through the University of Chicago's human subject recruitment system, SONA Systems. All subjects gave written, informed consent before participating. All of the methods performed in the study were in accordance with relevant safety and ethics guidelines for human subject research and were approved by the Social and Behavioral Sciences Institutional Review Board at the University of Chicago (IRB19-1302). This study was not a clinical trial.

2.2.3 Experimental design

Subjects completed three blocks of trials: a pretest block (30 trials), a stimulation block (250 trials), and a post-test block (30 trials). After initiating each trial, subjects waited for a visual indicator to cue their movements (Fig. 2.1b). After the visual indicator was triggered (2-4 s, uniformly distributed after trial start), subjects attempted to press a button (on a Cedrus RB-620 button box) as quickly as possible. After the button was pressed, the trial ended.

During the stimulation block, however, EMS was applied to the forearm after the cue to move, with the aim of preempting subjects' self-caused movement with an EMS-caused movement (Fig. 2.1c). After each trial, subjects were asked to report whether they caused the movement (agency) or the EMS caused the movement (nonagency).

Stimulation timing was adjusted on a trial-by-trial basis using a Bayesian optimization procedure designed to apply EMS as close as possible to the stimulation latency at which subjects would report agency with 50% probability (Figs. 2.1c, 2.2). Specifically, after each trial, we fit a Bayesian logistic regression predicting the probability of SoA from stimulation latency with a log-normal prior on the 50% threshold, centered 40 ms before subject's mean reaction time observed during the pretest block and a log-normal (and thus constrained to be positive) slope prior. This 40 ms prior on preemptive timing was based on that reported in previous work (Kasahara et al., 2019, 2021). Each trial's stimulation latency was drawn from the posterior distribution (truncated between 50 and 600 ms post-cue) of the 50% threshold in the logistic function fit after the previous trial. If the subject pressed the button before stimulation was delivered, stimulation occurred immediately on the button press so that the onset of the electrical stimulation, which causes a perceptible though painless tingling sensation on the skin, was always temporally confusable with that of the movement. However, because of the speed at which the optimization procedure converges to reliably preempt subjects' movements, such instances were quite rare (see Subsection 2.3.1).

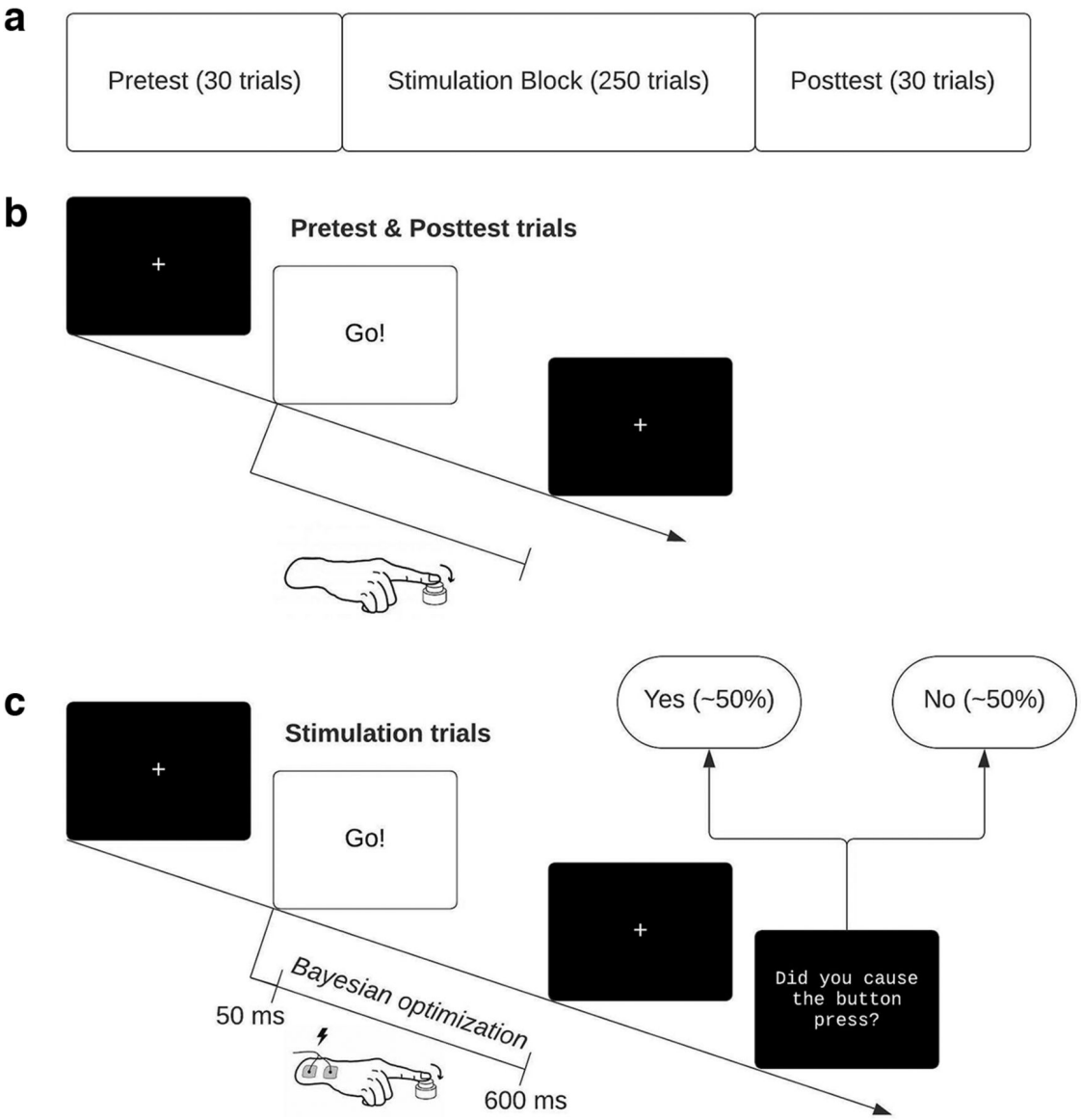

Figure 2.1: Task design. (A) Full experiment consists of a short pretest to gauge subjects' reaction times, a stimulation block, and a post-test block to ensure that true reaction times did not change dramatically over the course of the experiment. (B) Trials follow a typical cue-response reaction time paradigm, in which subjects are asked to press a button as quickly as possible following a cue to move. (C) In the stimulation block, subjects still attempt the reaction time task, but their natural movements can be preempted by muscle stimulation. After each trial, subjects guess whether the muscle movement resulting in the button press was self-caused or caused by muscle stimulation. Responses are used to tune the timing of muscle stimulation to a latency between 50 and 600 ms at which $\sim$50% of trials are perceived as self-caused via Bayesian optimization (as shown in Fig. 2.2).

Visual stimulus presentation was implemented using *PsychoPy* (Peirce, 2007) and Bayesian optimization using the Python-embedded probabilistic programming language *Pyro* (Bingham et al., 2019). All code has been made available (see Data and code availability).

2.2.4 Statistical analysis

Manipulation checks and outlier removal

First, outlier removal was applied to remove trials in which muscle movement was not caused by electrical stimulation. Thus, we removed trials in which (a) subjects pressed the button before stimulation, (b) the recorded response time was outside the stimulation time window (i.e., >600 ms), or (c) the lag between the EMS pulse and the corresponding button press fell outside of the middle 95% of the best fit log-normal distribution, indicating ineffective stimulation or the subjects' endogenous movement coinciding with the EMS-caused movement. These steps removed an average of 30.8 trials per subject, after measured "reaction time" (button press) is a roughly linear function of the stimulation latency (Fig. 2.3b).

To assess whether we were truly preempting subjects' movements, we then fit a Bayesian multilevel model to the trial-by-trial reaction times in each experiment block with the *Bambi* package (Capretto et al., 2022) using the conservative default priors (Westfall, 2017). If we were preempting subjects' movements, then reaction times in the stimulation block should be faster than in both the pretest or the post-test block.

As a final manipulation check, we then assessed whether stimulation latencies differed systematically between agency and nonagency trials. While prior work has shown that agency judgments vary as a function of stimulation latency (Kasahara et al., 2019, 2021; Tajima et al., 2022), the aim of our Bayesian optimization procedure was to minimize this confound by maximizing the overlap in stimulation latencies across agency and nonagency trials. Consequently, we fit a logistic regression predicting agency judgments from stimulation latency (with a random effect per subject, as in our EEG decoding analysis) to test whether

any residual relationship between the two is strong enough to drive our EEG results.

Linear EEG decoding

After preprocessing of the EEG signal (described in EEG acquisition and preprocessing), we assess the temporal dynamics of patterns which differentiate agency and nonagency trials using the temporal generalization method (King and Dehaene, 2014). In this approach, a linear classifier is trained on the pattern of voltages at each time point (using a training set), and then its classification performance is quantified at every other time point (using a test set), yielding information about both the occurrence and duration of neural patterns which predict the outcome of interest.

In our case, we use a logistic regression (fit using generalized estimating equations to account for subject-level random effects) as a linear classifier (Liang and Zeger, 1986; Seabold and Perktold, 2010) to predict agency judgments, and we quantify classification performance using the area under the receiver-operator curve (ROC-AUC), a nonparametric, criterion-free metric of class separation. By "criterion-free," we mean that, unlike metrics such as accuracy which depend on a particular decision boundary, ROC-AUC reflects the trade-off between false positives and false negatives across all possible decision boundaries; because of its weak assumptions, approximately normal distribution under the null hypotheses, and robustness to class imbalances, ROC-AUC is often recommended for multivariate pattern analyses of the EEG (King and Dehaene, 2014). Classification performance is calculated only on hold-out subjects (i.e., subjects not seen during classifier training), in a stratified 10-fold cross-validation scheme repeated 10 times. Cross-validated ROC-AUC scores are compared with chance performance (ROC-AUC = 0.5) using a one-sample t test with a variance correction to account for nonindependence between ROC-AUC values computed across cross-validation splits (Nadeau and Bengio, 1999).

This results in an $n_{\text{times}} \times n_{\text{times}}$ matrix M, where M_{ij} is the performance of the classifier

trained at time i evaluated at time j, as well as a p value for each (i, j) pair. The "shape" of above-chance decoding performances can then be interpreted as providing information as to the temporal characteristic of predictive patterns of neural activity. For instance, if a pattern is predictive only on the diagonal ($i = j$), that pattern is transient. On the other hand, if classification performance remains above-chance off-diagonal ($j > i$), then one can conclude the same pattern persists (and continues to predict SoA) across time. However, such conclusions are only licensed if one corrects for multiple comparisons using a method that allows inference about the "shape" of an effect, which common cluster-based corrections in the EEG literature do not (Sassenhagen and Draschkow, 2019). We use All-Resolutions Inference (Rosenblatt et al., 2018), which can compute simultaneous lower bounds on the true positive proportion in each cluster across all $n_{\text{times}} \times n_{\text{times}}$ possible clustering thresholds. This approach conveys uncertainty about the localization of true effects within clusters. For instance, if the proportion of true positives in a cluster is low, then one can conclude there are true positive effects within the cluster but it is unclear precisely where; conversely, if the proportion is high (e.g., >95%), then the localization is quite certain.

Complexity-based EEG decoding

In this analysis, we assessed whether certain complexity measures of EEG response to stimulation (the fractal dimension and the Hurst exponent) could predict agency judgments. These metrics measure nonlinear properties of time series which can be used to inform qualitative claims about those time series' underlying dynamics.

The fractal dimension, which we estimate using Higuchi's algorithm (Higuchi, 1988), is a measure of the complexity or "roughness" of a time series (or of its underlying dynamical attractor). The fractal dimension of both the background EEG and the EEG response to perturbation is highly predictive of states of consciousness (Kesić and Spasić, 2016; Ruiz de Miras et al., 2019), consistent with some accounts of conscious awareness (Oizumi et al.,

2014). Some preliminary evidence suggests that fractal dimension is higher for conscious percepts that are internally generated (e.g., mind wandering), making it a reasonable candidate predictor for SoA (Ibáñez-Molina and Iglesias-Parro, 2014). However, the interpretation of the fractal dimension on its own it ambiguous; it can be interpreted as reflecting how "self-similar" or "scale-free" a time series is, or alternatively, as reflecting the local complexity of its dynamics. These interpretations can be disambiguated in the context of the Hurst exponent.

The Hurst exponent, which we estimate using rescaled range analysis, is a measure of long-range temporal dependencies in a time series (Qian and Rasheed, 2004). In the cognitive neuroscience literature, these long-range dependencies have been argued to reflect how much local events, such as an external input, can alter the course of a neural system, assuming that events that substantially impact the system should have consequences which persist in time (Churchill et al., 2016; Kardan et al., 2020; Zhuang et al., 2022). Hurst is notably suppressed in those suffering from psychiatric conditions associated with an impaired SoA (Sokunbi et al., 2014; Stier et al., 2021).

If a time series is strictly self-similar, then the fractal dimension D will be related to the Hurst exponent H by the deterministic relationship $D + H = 2$ (Gneiting and Schlather, 2004), but these metrics have been reported to diverge in EEG data despite the $1/f$ power spectrum of resting/background EEG, implying some degree of self-similarity in the signal (Martis et al., 2015). Unfortunately, estimates of the Hurst exponent computed from time series as small as our single-trials are known to be biased (Oliver and Ballester, 1998; Eke et al., 2000, 2002). While the bias of our Hurst estimates prevents us from testing the $D + H = 2$ relationship directly, within-subject variation in trial-by-trial Hurst exponents estimated from EEG has been shown to be sensitive to cognitive functions (Kardan et al., 2020). Thus, if the EEG time series is self-similar or scale-free, then should the fractal dimension increase with agency, Hurst should decrease and vice versa. However, if they both positively or both negatively covary with agency (or one covaries and not the other), then

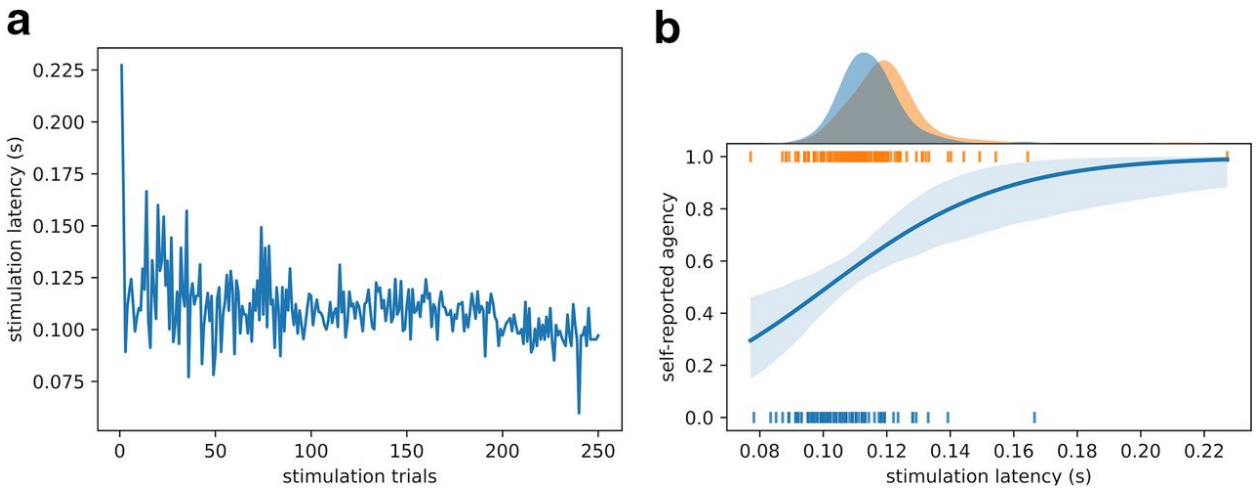

Figure 2.2: Trial-by-trial stimulation latency over the course of the stimulation block for a representative subject. (A) Stimulation latency hones in on a stable value over time, as a result of the Bayesian optimization procedure. (B) A logistic regression computed after the experiment shows that stimulation times are close to the retroactively estimated 50% threshold, although that threshold was not known in advance. The subject featured here is "sub-07" in the associated dataset.

the EEG response to stimulation is unlikely to be self-similar.

Since we perform this analysis at the electrode level (instead of training a single classifier on all electrodes), we apply a current source density transformation before computing per-electrode complexity measures to increase the interpretability of spatial information in the EEG signal (Kayser and Tenke, 2015). For each electrode and subject, then, we compute the ROC-AUC between both of these metrics and subjects' self-reported agency. Since there are no fit-to-the-data parameters in this analysis, no cross-validation scheme is necessary; a one-sample, two-sided t-test is used compare subject-level decoding performance to chance (ROC-AUC = 0.5).

2.2.5 Electrical muscle stimulation

Before the experiment began, two EMS electrodes were applied to the skin above the flexor digitorum profundus muscle on the right (dominant) forearm, which is an easily accessible muscle that moves the ring finger, which subjects used to press the button during the exper-

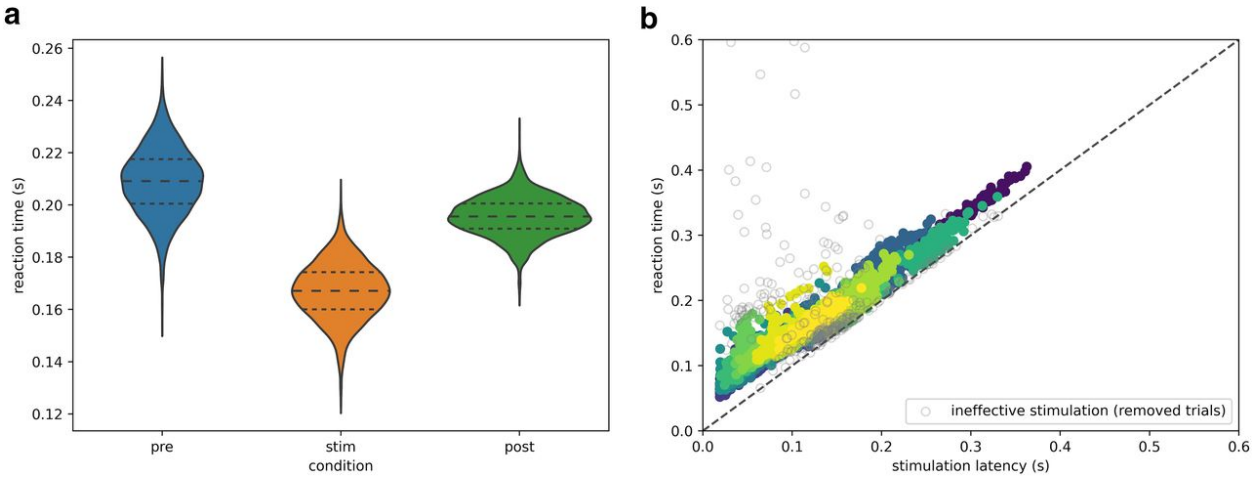

Figure 2.3: EMS consistently preempted subjects' volitional movements. (A) Posterior distributions of the mean reaction times in each condition show that EMS-induced muscle movements occur earlier than subjects' natural muscle movements. (B) After outlier removal, measured "reaction" times (shown for all trials and subjects) are a nearly linear function of the stimulation latency, indicating that movements in the remaining trials are, indeed, EMS-actuated.

iment. Stimulation was performed with a RehaStim 1 device (HASOMED). On each trial, muscle actuation consisted of a single, biphasic pulse of constant current stimulation lasting 900 microseconds (400 µs positive, 400 µs negative, separated by 100 µs).

Before beginning the experiment, we calibrated the stimulation amplitude to the minimum intensity required to reliably move the subject's finger. The calibration procedure was as follows: (1) The subject placed their ring finger on the button that would be used during the experiment and was instructed not to move their hand. (2) Starting at an intensity of 1 mA, we stimulated the subjects' arm 10 times. If $<$10 button presses were registered, then we iterated the intensity by 1 mA and repeated. (3) We stopped increasing the intensity on achieving 10 consecutive actuated button presses, or if a conservative safety limit of 25 mA was reached.

2.2.6 EEG acquisition and preprocessing

EEG was recorded with 64 active Ag/AgCl electrodes (actiCHamp, Brain Products) placed according to the extended 10–20 system. At the time of recording, the electrodes were referenced to Cz and sampled at 10,000 Hz. Two of the 64 electrodes (which would have been AF7 and AF8 on the typical actiCAP layout) were dropped below the left and right eyes so that they could later be rereferenced to become EOG channels. Experiment events were marked in the EEG recording using TTL triggers and later corrected with a photosensor (Brain Products) on the subjects' screen. The precise subject-specific positions of the 62 head electrodes were measured at the end of each recording using a CapTrak (Brain Products).

EEG was later preprocessed in Python using *MNE-Python* package (Gramfort et al., 2014). First, we fit a multitaper estimation of the sinusoidal components at the line noise frequency and its first two harmonics to partially attenuate electrical interference before interpolating the stimulus artifact. Then, the electrical artifact from the EMS pulse was removed by linearly interpolating over the interval starting 5 ms before and ending 10 ms after the event time stamp. Then, after interpolation, we applied an additional FIR notch filter at 60 Hz and its harmonics up to the intended upper passband edge (see below) to thoroughly clean the data of line noise, and then resampled the data to 5000 Hz to improve the speed of computation for subsequent preprocessing steps.

Next, we applied common preprocessing operations in adherence with the standardized PREP preprocessing pipeline for EEG data (Bigdely-Shamlo et al., 2015) using the implementation in the *PyPREP* package (Appelhoff et al., 2022). This pipeline robustly rereferences the EEG signal to the average of all electrodes and interpolates electrodes it determines have poor signal quality; see the PREP paper for a full description (Bigdely-Shamlo et al., 2015). A record of which channels were interpolated is available in subject-specific preprocessing reports (see Data and code availability).

Then, we filtered the data to the frequency band used for analysis. We used a single low cutoff of 1 Hz to remove low-frequency drift, but we used different high cutoff values for the different analysis described in Linear EEG decoding and Complexity-based EEG decoding. For linear decoding (see Linear EEG decoding), we used a 30 Hz high cutoff; this filter setting is common for the analysis of event-related potentials, as this level of temporal smoothing helps to align short neural events across subjects (Luck, 2014), and we posited that such smoothing would likely improve sensitivity for between-subject decoding as we use here. However, since fractal dimension is fundamentally a measure of signal roughness, which would be distorted by anything that would artificially smooth the signal, we used a more liberal 70 Hz high cutoff for the fractal analysis described in Complexity-based EEG decoding.

We then removed EOG contamination of the EEG signal. We decomposed to EEG data into 15 independent components (ICs) using the FastICA algorithm (Hyvarinen, 1999). Then, we correlated each IC with the EOG channels, z-scored the correlation coefficients, and deemed an IC to contain eye artifact if the absolute value of its z score exceeded 1.96. Those ICs were zeroed out to remove them from the original data. Plots of the scalp topographies of removed ICs for each subject can be found in their preprocessing reports (see Data and code availability).

Subsequently, we segmented the data into epochs starting 100 ms before the onset of stimulation and ending 500 ms after stimulation. We then estimated the peak-to-peak rejection threshold that would be used to identify trials containing unwanted artifacts using the *Autoreject* package (Jas et al., 2017), which estimates the optimal threshold as that which minimizes the fivefold cross-validated root-mean-squared difference between the mean of the training folds and the median of the testing fold, a robust proxy metric for signal-to-noise. The resulting per-subject rejection thresholds are recorded in each subjects' preprocessing report (see Subsection 2.2.7).

Since the visual evoked response to the movement cue is unlikely to be over by the time of stimulation, we attempted to remove the visual evoked response from our epoched data to minimize confounds. To do so, we computed evoked responses to both the visual and electrical stimuli simultaneously using a regression-based overlap correction on the continuous (nonepoched) data, excluding second-long chunks of the data in which peak-to-peak amplitude exceeds the rejection threshold (Smith and Kutas, 2015); conceptually, this is very similar to the way GLMs are used to deconvolve hemodynamic responses in fMRI. Then, the overlap-corrected visual evoked response was aligned with the epoched version of the data and subtracted out. Thus, the average visual response to the movement cue was removed from the stimulation-locked epochs. Subject-level evoked responses can be found in our open dataset and are visualized in the subject-specific preprocessing reports (see Subsection 2.2.7).

Finally, the rejection threshold was applied to the cleaned and overlap-corrected epochs, removing trials still contaminated by artifacts. The surviving epochs were down-sampled to twice their high-cutoff frequency for computational expediency and saved for further analysis. This epoched data are available in our open dataset, and subject-level trial yields are recorded in the accompanying quality check reports (see Subsection 2.2.7).

2.2.7 Data and code availability

Code for running the experiment can be found on GitHub (github.com/apex-lab/agency-experiment) and in a permanent archive on Zenodo (doi.org/10.5281/zenodo.7894011). Similarly, all data analysis code, including EEG preprocessing code, can be found at https://doi.org/10.5281/zenodo.8367570. All data, including both raw data, preprocessed derivatives, and postpreprocessing quality check reports for each subject, can be found on OpenNeuro (doi.org/10.18112/openneuro.ds004561.v1.0.0).

2.2.8 Statistical power

There is no widely agreed-on approach for estimating the statistical power for detecting novel EEG effects, in which the spatiotemporal distribution of the effect is unknown *a priori*, as we recently reviewed (Veillette et al., 2023a). Statistical power for EEG effects depends not just on the number of subjects but also on the number of trials, and how these two design considerations interact to affect power seems to differ between components of the EEG response (Boudewyn et al., 2018; Hall et al., 2023; Jensen and MacDonald, 2023). However, statistical power for well-known EEG effects has been studied using a recently introduced Monte Carlo simulation approach (Boudewyn et al., 2018), and it is worth considering how well our study is powered for detecting effects reported in the literature. While we and others have found, using such a simulation-based approach, that a relatively small number of subjects and trials achieves very high statistical power for detecting the presence of seven endogenous EEG-evoked response effects (Jensen and MacDonald, 2023; Veillette et al., 2023a), our main study result, that which differs from previous EEG studies of SoA, concerns an early (<200 ms) effect, and such effects usually reflect amplitude changes in exogenous response components present in both conditions rather than the presence or absence of an endogenous component. This more realistic case has been studied for three early evoked response components (Hall et al., 2023). Closest to our sample size, Hall et al. (2023) reported that a within-subject design with a sample of 25 subjects, each having 120 trials per condition, achieves a power of at least 0.8 for detecting a 1.4 µV amplitude difference in the N1 component (in the window of 84-124 ms), a 1.3 µV difference in the Tb component (124-164 ms), and a 1.7 µV difference in the P2 component (151-191 ms) with a significance level of 0.05. Based on this comparison, we would expect our linear classifiers to be sensitive (i.e., with power of ∼0.8) to amplitude differences on the order of ∼1.5 µV.

2.3 Results

2.3.1 Bayesian optimization effectively controls the proportion of trials perceived as self-caused

The Bayesian optimization procedure resulted in trial-by-trial stimulation latencies honing in on some threshold estimate throughout the stimulation block. A representative time course is shown in Figure 2.2.

After removing trials in which stimulation failed to produce a muscle movement (and therefore the "reaction" time was not a function of stimulation latency), our multilevel model of the recorded reaction times estimated a 99.9% posterior probability that button presses occurred earlier in the stimulation block than in either of the other blocks. In particular, we estimate that "reaction" times resulting from EMS-actuated movements were between 17.5 and 65.0 ms faster than true reaction times in the first (pre) block with 95% probability, and between 13.8 and 43.9 ms faster than those in the final (post) block. A nominal speedup between the pre and post blocks was observed with 90.6% probability (95% HDI: [-6.7 ms, 32.8 ms]), suggesting that subjects may have improved their reaction times by the end of the task, but not enough to account for the much lower reaction times in the stimulation block. Posterior distributions for the (group) mean response times in each condition are shown in Figure 2.3. Together with the near linear relationship between stimulation latency and reaction time, we can conclude that movements were usually caused by muscle stimulation rather than the subject, effectively preempting subjects' volitional movements.

While it is evident that muscle movements in the stimulation block (after outlier removal) were overwhelmingly caused by EMS rather than by the subject, subjects still reported that they caused roughly half of the movements. Overall, after outlier removal (Fig. 2.3), 51.98% of all trials across all subjects were judged as self-caused. On average, subjects reported that they caused 50.99% (SD: 14%) of movements. In other words, the Bayesian optimization

procedure was effective at controlling the proportion of trials in which movements were experienced as self-caused, generating an ∼50–50 split of agency versus nonagency trials.

While it is understood that agency judgments in this task paradigm vary as a function of the stimulation latency (Kasahara et al., 2019, 2021; Tajima et al., 2022), our Bayesian optimization procedure converges to a narrow latency range around the 50% agency threshold quickly enough to attenuate this confound. A logistic regression predicting agency judgments from stimulation latency (with a subject-level random effect), notably the same approach we use to predict agency judgments from the EEG signal, fails to find a statistically significant relationship between the two ($\beta = 0.95$, 95% CI: [-0.91, 2.82], $p = 0.315$). Thus, any residual relationship between stimulation latency and SoA is unlikely to explain our EEG findings (see below).

2.3.2 Distinct early and late neural processes predict agency judgments

Our linear decoding procedure showed above-chance decoding performance across subjects, reaching up to ROC-AUC = 0.587; thus, the patterns which we report predict agency judgments generalize across individuals. While we report the true-positive proportion within clusters across all clustering thresholds (Fig. 2.4b), we will focus primarily on the clusters in which the true positive proportion exceeds 95%, since these clusters are where we are sufficiently certain about the localization of the effect (Rosenblatt et al., 2018). The grand-average EEG-evoked response to muscle stimulation is provided, for visualization only, in Figure 2.5; this may be useful context when considering predictive topographies (Fig. 2.6).

The earliest such cluster occurs 83 ms after the onset of muscle stimulation (adjusted threshold: $p < 4.5\times$ 10–6). This is substantially earlier than previous studies have localized the earliest predictors of agency judgments (see Discussion), which may reflect a distinct role of low-level sensorimotor processes in agency judgments pertaining to the musculature itself, but less so to downstream sensory consequences of action. When comparing the patterns

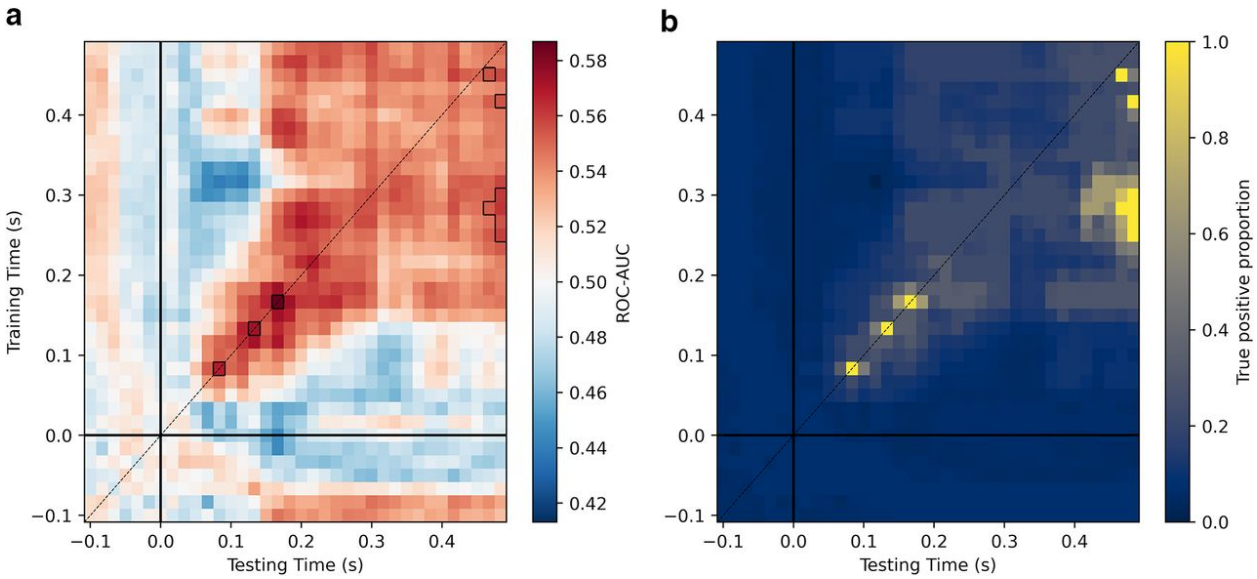

Figure 2.4: Temporal generalization of neural patterns predicting SoA. (A) Classification performance (ROC-AUC) for decoding subjects' judgment of agency for individual muscle movements, cross-validated across subjects and across time. Results are shown for all (train-time, test-time) pairs to visualize the temporal dynamics of patterns that predict SoA. Above-chance decoding only near the diagonal reflects neural patterns that predict agency only transiently, whereas above-chance decoding far off-diagonal reflects patterns that are sustained over time. Thus, patterns predicting agency appear to transition from transient to sustained dynamics at ∼170 ms following stimulation. (B) Lower bounds on the true-positive proportion within clusters, computed across all clustering thresholds. The value represented at each (train-time, test-time) pair is the highest true-positive proportion of any cluster in which that pair is included; thus, larger values reflect greater certainty in the localization of effects.

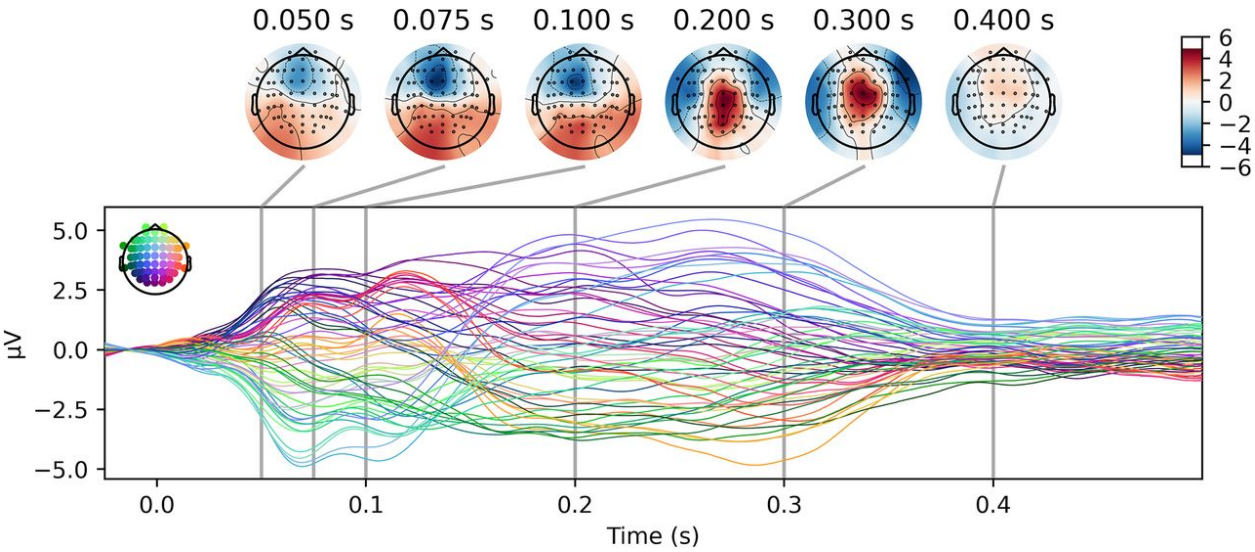

Figure 2.5: The depicted waveform was computed by averaging the preprocessed (1-30 Hz filtered) data across stimulation trials within each subject, and then averaging the resulting subject-level EEG responses to obtain a group-level average.

our decoding model selects for (Fig. 2.6) to the average evoked response (Fig. 2.5), one notes that the polarity of the pattern that predicts SoA is opposite the average response, indicating that the classifier would predict a self-agency judgment as the result when the sensory response is suppressed, a finding consistent with sensory attenuation (Voss et al., 2006). Classifiers trained at earlier times do not generalize to predict SoA at later times (Fig. 2.4), indicating that early prediction likely reflects a sequential chain of transient representations during sensorimotor processing (King and Dehaene, 2014). Later in the epoch, however, the temporal dynamics of the predictive patterns change to reflect a single, sustained neural signature that predicts SoA starting by at least 250 ms after stimulation and persisting at least until the end of the epoch ($p < 0.003$).

2.3.3 Fractal complexity of brain activity predicts agency judgments

Notably, trial-by-trial fractal dimension predicted SoA at almost every electrode (Fig. 2.7), reaching an ROC-AUC of 0.614 at electrode C1 (adjusted threshold: $p < 0.027$), even after the current source density transformation of the EEG signal was applied to attenuate the

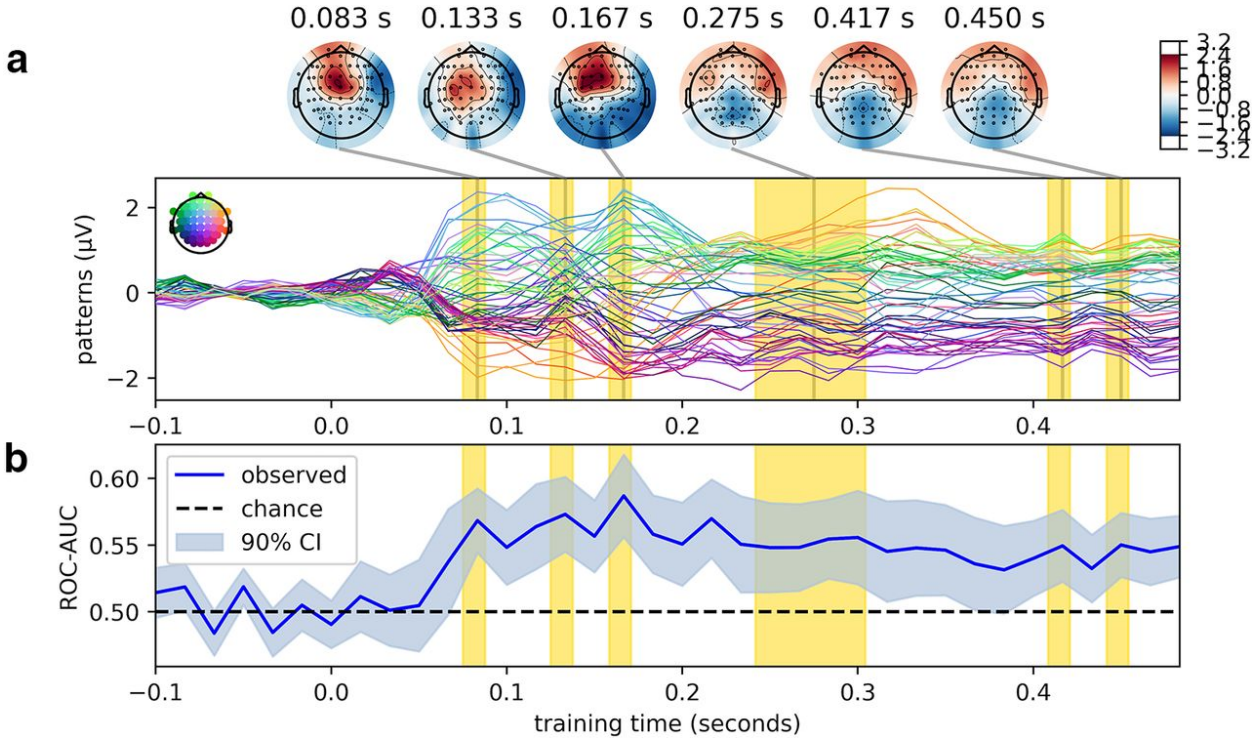

Figure 2.6: Voltage patterns that predict SoA. (A) The EEG topographies that the linear classifiers trained at each time point select for, reconstructed by inverting the trained classifier parameters using Haufe's trick (Haufe et al., 2014). (B) The decoding performance when testing at each train time (identical to the values on the diagonal in Fig. 2.4a). Training times are highlighted in yellow if included in a cluster with true-positive proportion > 0.95 at any test time (Fig. 2.4b).

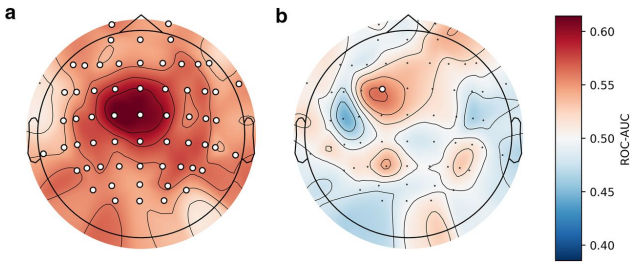

Figure 2.7: Classification performance for predicting trial-by-trial SoA from single-electrode fractal metrics. (A) Classification performance for Higuchi fractal dimension. (B) Classification performance for the Hurst exponent. Electrodes included in clusters in which the true positive proportion exceeds 95% are marked with white.

effects of volume conduction (see Materials and Methods). This suggest that the (local) complexity of the brain activity is increased uniformly throughout cortex following muscle movement when that movement is perceived as self-caused (compared with when it is not perceived as self-caused). This is consistent with the previous observation that neural activity corresponding to self-generated percepts has a higher fractal dimension (Ibáñez-Molina and Iglesias-Parro, 2014).

On the other hand, the Hurst exponent only predicted SoA at a single electrode at position FC1 (ROC-AUC = 0.559, p = 0.0006), located above cortical regions involved in motor control and planning, contralateral to the arm in which stimulation occurred (though we did not vary the arm used for stimulation, so we would caution against interpreting this as a strictly contralateral effect, though it is suggestive). This finding suggests a much more selective modulation of long-range temporal dependencies, such that the activity of specific frontocentral cortical regions becomes globally less to local perturbations in the absence of SoA; or, conversely, frontocentral areas are more sensitive to inputs in the presence of SoA. Notably, since the Hurst exponent (albeit only in one electrode) and fractal dimension both positively covary with agency, defying the strictly inverse relationship they would show in a strictly self-similar time series (see Complexity-based EEG decoding), the EEG response to muscle movement appears to depart from (full) scale-freeness, at least over FC1. This divergence would allow the local complexity of and the temporal persistence of perturbations

to neural activity to be modulated independently (see Discussion below).

2.4 Discussion

Our findings advance our understanding of how the SoA is generated in the brain, with important implications for the relationship between conscious self-awareness and unconscious self-referential processing. In particular, the time course neural activity predicting SoA in response to muscle stimulation is more consistent with classical sensorimotor monitoring accounts (Wolpert et al., 1995; Blakemore et al., 2000) than previous studies have shown comparing the neural responses to self- and other-caused tones (Kühn et al., 2011; Timm et al., 2016). While results still leave room for common downstream correlates of agency, they suggest that early responses differentiating self and other may be more modality-specific than previously thought.

In the comparator model of SoA, originally imported from the motor control literature (Wolpert et al., 1995), sensations are compared with the intended or predicted sensory consequences of actions, and then congruent feedback is deemed self-caused and incongruent feedback externally caused (Feinberg, 1978; Frith, 1987; Gallagher, 2000). Since it is well documented that early sensory responses, especially those that are predictable, are attenuated during movement (Blakemore et al., 2000), it seemed plausible that the same machinery could parsimoniously account for conscious self-other discrimination. While this simple model is still the basis of most modern accounts of SoA, it is now understood that the mechanisms of conscious SoA diverge from low-level sensorimotor monitoring (Synofzik et al., 2008; Frith, 2012; Press et al., 2023).

To this effect, recent studies using typical paradigms, which probe the perception of causality between a button press and subsequent tone (i.e., "broad" SoA over action outcomes), have failed to find a relationship between the neural processes which would be affected by low-level sensorimotor monitoring; that is, the early, preconscious (<200 ms) re-

sponse to sensory stimulation and conscious SoA (Voss et al., 2006; Kühn et al., 2011; Ohata et al., 2020). Timm et al. (2016) report a full dissociation, showing that comparator-model-like suppression of early responses to self-caused sensation occurs in both the presence and absence of SoA. Since decades of research tell us early (<200 ms) sensory responses reflect preconscious, rather than conscious, processing of the sensory stimulus (Libet et al., 1967; Sergent et al., 2005; Dehaene and Changeux, 2011), these findings have been interpreted as meaning that temporally early "exogeneous" neural responses (i.e., those that are a direct consequence sensory input) do not inform agency judgments, but later "endogenous" neural responses (e.g., P3 component) associated with conscious attention do (Kühn et al., 2011). None of these authors argue against the general idea of a comparator, but rather suggest that the comparison takes place at a higher level of abstraction than in the low-level sensorimotor monitoring used to guide motor learning (Wolpert et al., 2011).

In contrast, we find patterns in the early sensorineural response to stimulation predicts SoA, even when that sensation was not actually self-caused, as we exclusively analyzed trials in which movements were caused by EMS. The critical difference is that we measured the neural response to muscle stimulation, and subjects made agency judgments about the muscle movement itself rather than a downstream consequence of movement. Thus, the mechanisms that give rise to narrow SoA over the musculature may overlap with basic sensorimotor processing more than those mechanisms that give rise to SoA over action outcomes more far removed from a subject's motor intention (Charalampaki et al., 2022). Previous work manipulating bodily agency by altering the visual feedback from movement (leaving somatic feedback channels intact) has primarily used fMRI (Tsakiris et al., 2010; Abdulkarim et al., 2023) or EEG methods lacking the temporal resolution of the present approach (Kang et al., 2015); consequently, it is not totally clear whether our very early (preconscious) decoding results differ from previous findings merely because of our focus on SoA over body movements or because we additionally perturbed somatic (not just visual) feedback channels. Regardless,

our data support the view that the earliest (preconscious) correlates of conscious SoA may differ based on context (i.e., what is one being asked to make a judgment about?), modality (e.g., proprioceptive or auditory), or level of abstraction.

However, it is worth noting that the earliest neural correlates of agency are not the end of the story. Indeed, the comparator model for SoA has largely been usurped by dual-process models in which the outcome of an initial comparator process is integrated with prospective, prior information to produce a final agency judgment (Synofzik et al., 2008; Haggard, 2017; Legaspi and Toyoizumi, 2019), and there is no clear theoretical for why or how multiple comparator processes taking place at multiple levels of abstraction may not be integrated into a single agency judgment. Indeed, the shift we observe from transient to sustained patterns of neural activity predicting agency is quite consistent with that predicted by dual-process models of action processing (Del Cul et al., 2009; Charles et al., 2014). Specifically, the sustained nature of the predictive voltage patterns is consistent with a previously observed signature of high-level novelty/error detection that has been argued to require conscious awareness (Dehaene and King, 2016) and previously proposed to inform agency judgments (Kühn et al., 2011). An intriguing possibility then, which hybridizes the competing views proposed in the Introduction, is that preconscious ($\sim<$200 ms) predictors of SoA judgments will be context-specific, but postconsciousness "neural correlates of self-awareness" integrate across modality-specific comparators. We do not manipulate awareness of action and outcomes here, so it is up to future work to test this hypothesis directly. Such investigations, which can compare SoA over actions with SoA over those actions' downstream outcomes, are made possible by extending the paradigm we introduce here.

Further, both the fractal dimension, a measure of local signal complexity or "roughness," and the Hurst exponent, a global measure of long-range correlation in a signal, indicative of how long a perturbation (e.g., sensory input) in the measured system would persist in time, were able to classify trial-by-trial SoA. However, the Hurst exponent was only predictive of

SoA in a single frontocentral electrode, whereas fractal dimension was robustly predictive across the whole scalp. Both of these measures are often interpreted as reflecting a self-similarity or scale-free property of a time series, often appealing to theories of self-organized criticality as an explanatory framework (Churchill et al., 2016; Kardan et al., 2020; Zhuang et al., 2022). Indeed, the self-similarity interpretation has been invoked in explaining why the fractal dimension of neural activity corresponding to self-generated percepts is higher than that to external stimuli (Ibáñez-Molina and Iglesias-Parro, 2014). In a truly self-similar time series, however, fractal dimension and the Hurst exponent are strictly inversely related (Gneiting and Schlather, 2004); in contrast, both values positively covaried with SoA in the electrode in which we find Hurst was predictive. This finding suggests the neural response to muscle movement (as reflected in EEG) is not strictly self-similar, and so its complexity and sensitivity to perturbation can vary independently. While admittedly quite speculative, this observation may be interpreted as having functional importance, allowing sensorimotor cortical regions (which could possibly account for the frontocentral Hurst effect) to selectively modulate sensitivity to input, while overall cortex shows higher signal complexity with SoA.

In conclusion, while SoA has become a topic of increased attention in recent decades, most research in the area has focused on the experience of agency over downstream consequences of one's actions as they affect the external world rather than the more basal experience of directing one's own muscles (Haggard, 2008, 2017). We introduce the use of human-in-the-loop Bayesian optimization, in combination with EMS, to experimentally manipulate the subjective experience of controlling the musculature. As we showcase here, this approach enables novel behavioral and neuroimaging investigations into the substrate of embodied self-awareness. Our results provide confirmatory evidence for the predictive relationship between low-level sensorimotor processes and SoA for muscle movements, which seems not to hold for the sensory response to action consequences (Dewey and Knoblich, 2014; Timm et al., 2016). While our findings suggest that early neural correlates of SoA may differ by context

and modality, the transition from transient to sustained neural patterns that predict SoA in our data suggest at least two distinct neural processes contributing to agency judgments, as posited by dual-process theories of action selection and monitoring (Del Cul et al., 2009). This leaves open the possibility that modality-specific, preconscious predictors of SoA are still integrated into a single agency judgment downstream. Such a possibility could explain how information from multiple scales of biological organization are integrated into a unified experience of self, even if the mechanism of self-other differentiation differs across scales. We suggest that this hypothesis is a fruitful avenue of research for the emerging science of self-awareness.

CHAPTER 3

METACOGNITION BRIDGES EXPERIENCES AND BELIEFS IN SUBJECTIVE AGENCY

3.1 Introduction

Sense of agency (SoA) is the feeling or belief that one controls one's own actions and, through those actions, can influence events in the world. We experience a feeling of "I did that" as we intentionally take action. This phenomenological SoA, together with body ownership, has been argued by philosophers, psychologists, and neuroscientists alike to be the most basic building block of a minimal conscious self-awareness (Gallagher, 2000; Tsakiris et al., 2006). However, beyond the "minimal self," SoA is also discussed as a high-level belief about one's level of control incorporated into the "narrative self" that is sustained over time (Dennett, 1993; Gallagher, 2000). An important aspect of this idea is that self-reported beliefs about one's own agency appear to be relatively stable over time, constituting a trait-level phenomenon (Tapal et al., 2017). It is assumed by some that the more elaborated narrative self is, in some way, constructed from our moment-to-moment experience of minimal selfhood, but the operating characteristics of this putative integration have been left vague (Gallagher, 2000). Are sustained beliefs about the self merely the sum of individual experiences, or is there more to the construction of declarative beliefs about one's agency? While it is common for studies in the literature to refer to "the" sense of agency, SoA has been discussed as a heterogenous psychological phenomenon for decades (Charalampaki et al., 2022; David, 2012; Gallagher, 2012; Pacherie, 2007). As such, researchers have employed domain-specific measures to study SoA and related processes at various scales of psychological-behavioral organization, ranging from in-the-moment detection of sensorimotor contingencies to declarative beliefs about one's status as an agent that are sustained over longer periods of time. How measures of SoA across different domains relate to one another remains unclear; indeed,

some researchers have argued that the concept of agency as it pertains to the narrative self is totally independent from that of the sensorimotor self (Jenkins, 2001).

At a sensorimotor level, humans and other organisms are constantly engaged in a process of distinguishing between self-caused and externally-caused sensory stimuli. A basic expression of this distinction is the neural suppression of predictable sensations resulting from voluntary movement (Frith, 1987) . That is, sensations are perceived as self-caused when they are predictable from intended actions and therefore attenuated, but whether one experiences agency over a particular sensation further depends upon other cues such as prospective movement intentions (Frith, 2012; Haggard, 2017) or contextual information (Desantis et al., 2011; Synofzik et al., 2008). In the same manner that detection thresholds vary across participants in a variety of sensory domains (Hirsh and Watson, 1996; Stevens and O'Connell, 1991), the degree to which the sensory consequence of an action must match the predicted consequence before control is reliably detected may vary from subject to subject. One's sensitivity to cues to their own control has been measured using control discrimination tasks, in which subjects identify which of two stimuli they are able to influence with their movements (Wang et al., 2020; Wen and Haggard, 2018, 2020). Eliciting uncertainty ratings about agency judgments during control discrimination tasks also allows one to measure metacognitive sensitivity concerning such judgments. While, depending on a number of contextual factors, judgements of agency may be influenced by metacognitive processes (Chambon et al., 2014), a combination of empirical findings and computational modelling has been used to convincingly argue that agency judgements – once considered inherently metacognitive – need not depend on metacognitive processing (Constant et al., 2022; Wen et al., 2023). Consequently, role of metacognition in the awareness of agency remains a topic of intense debate.

In addition to whether one feels agency over an action or outcome (i.e. an agency judgement), one can also discuss phenomenal experiences that either result from or tend to ac-

company judgements of agency. A basic finding is that, when one experiences agency over an action-outcome pairing (say, a button press with a resulting tone), the perceived time of the action is shifted toward the time of the outcome and vice versa, putatively intentionally binding the two into a single event according to the common theoretical interpretation (Haggard et al., 2002). While often used as an implicit measure of agency, the magnitude of the intentional binding effect is typically increased when subjects produce intentional actions but does not depend on explicit consideration of agency per se (Lush et al., 2017). Thus, the measure corresponds to a change in conscious experience that could co-occur with instances of control regardless of whether a subject attends to such instances.

At the level of the narrative self, explicit declarative beliefs about one's SoA can be measured psychometrically using the Sense of Agency Scale (SoAS) (Tapal et al., 2017). Two factors, termed "positive" SoA (SoPA) and "negative" SoA (SoNA) can be derived from subject responses to the scale items. The stability of this factor structure has been confirmed and replicated, and these factors can be differentiated from other related constructs such as self-efficacy beliefs and free will beliefs, and there is high test–retest reliability that can be seen even when separated by months (Hurault et al., 2020; Tapal et al., 2017). Moreover, the measured factors predict obsessive–compulsive symptoms and differ between patients with psychosis and healthy controls, which has suggested that these factors meaningfully, though necessarily crudely, quantify clinically important differences in subjective experience (Kruse et al., 2022; Tapal et al., 2017).

The present study, then, aims to assess the degree to which declarative beliefs about agency are predicted by moment-to-moment agency judgements in our sensorimotor interactions. We use validated online tasks to measure, for each subject, sensitivity to sensory evidence of control during agency judgements and the accuracy of metacognitive monitoring of those judgements (Wang et al., 2020) as well as the magnitude of the intentional binding effect as an index of how deeply inferences of agency affect perceptual awareness (Galang

et al., 2021). We then estimate the extent to which individual differences in these indices of moment-to-moment SoA predict beliefs about SoA measured by the SoAS. Our results, accordingly, inform our understanding of how SoA at the sensorimotor level relates to SoA at the level of declarative beliefs.

3.2 Methods

3.2.1 Subject recruitment and ethics

200 subjects were recruited online from across the United States using Prolific (prolific.co), the behavioral task was hosted on Pavlovia (pavlovia.org), and all experiment code was written in JavaScript using the jsPsych library (de Leeuw, 2015). Subjects in Prolific's recruitment pool were only allowed to participate if 95% of their previous submissions on the site had been approved and they were using the Windows operating system. 5 subjects' data were lost due to technical error, resulting in n = 195. Subjects' data were excluded from analysis of particular tasks or scales (not removed from analyses of other tasks) if they failed to pass exclusion criteria/attention checks for that task, such as failing an unacceptably high proportion of trivially easy "catch" trials, or if they had partially missing data; that is, exclusion criteria were applied to each task's data separately, and each pairwise statistic comparing measures across tasks used the highest amount of usable data to minimize information loss. Please see task descriptions below for exact exclusion criteria and counts of subjects removed from each task. Subjects were 56.4% male, with mean age 39.8 (SD=12.7) and a median age of 37.0 years. Sample size was determined arbitrarily.

All subjects gave written, informed consent before participating. All of the methods performed in the study were in accordance with relevant safety and ethics guidelines for human subject research and were approved by the Social and Behavioral Sciences Institutional Review Board at the University of Chicago (IRB21-1458). This study was not a clinical trial.

3.2.2 Selection and summary of sensorimotor measures

Since ensuring acceptable stimulus timing for online experiments is inherently difficult, we opted to use only experimental tasks with previously validated JavaScript implementations to obtain agency-related measures in a sensorimotor setting. To limit our own analytic flexibility, and thus avoid biasing results, we initially calculated only – and all of, to avoid selection bias – the measures reported in these previous validation papers.

From the intentional binding paradigm, we report action binding (the amount the perceived timing of an action, i.e. keypress, is shifted toward its sensory consequence) and outcome binding (the amount the perceived time of the sensory consequence is reciprocally shifted toward the precipitating action), as reported in the validation paper for the task implementation (Galang et al., 2021). While within-subject differences in these intentional binding effects are frequently used as an implicit measure of SoA (Haggard, 2017), some evidence also suggests that binding strength correlates with the sensitivity of explicit reports of agency to experimental manipulations (Imaizumi and Tanno, 2019). Given, however, that intentional binding is also known to dissociate from explicit agency judgments (Suzuki et al., 2019), the present authors interpret the intentional binding effect at face value; that is, intentional binding effects are interpreted here as a phenomenal experience that often co-occurs with intentional action but it not contingent on subjects making an explicit judgment of agency (Lush et al., 2017, 2019).

In the control detection task, we quantify task performance using the control detection threshold (defined as the objective level of control at which subjects can identify that their mouse movements control a sensory stimulus with 75% accuracy), which reflects subjects' sensitivity to sensorimotor evidence of control. If two subjects encounter the same distribution of sensory evidence of control, the subject with a lower threshold (i.e. the subject with higher control sensitivity) would make positive agency judgments more often. Of course, not all people do encounter the same distribution of evidence of control – someone with a

motor deficit, for instance, may simply encounter fewer genuine instances of control – but this measure may be nonetheless thought of as an imperfect proxy for the frequency with which subjects make positive agency judgments based on sensorimotor evidence. This measure was found to have high test–retest reliability in the validation paper associated with the implementation of the control detection task we use (Wang et al., 2020). Additionally, we ask subjects to rate their uncertainty about their agency judgment after each trial, and we quantify the accuracy of their uncertainty judgments – type II error rate – using meta-d'. In an ideal observer under signal detection theory, first order d' (subjects' sensitivity to control in signal-to-noise units) and type II error rates are deterministically related, so meta-d' is defined as the d' one would expect to have produced the observed type II error rates if subjects had complete metacognitive access to the sensory evidence underlying their first order judgment (Fleming and Lau, 2014). Thus, meta-d' is theoretically interpreted as the amount of sensory evidence (again in signal-to-noise units) from their first order judgement to which subjects still have access when reporting their uncertainty; in our case, as we use an adaptive staircase to find subjects' control detection threshold as defined above, meta-d' specifically refers to the amount of sensory evidence of control to which they have metacognitive access at their detection threshold. However, the measure can be more agnostically interpreted as a measure of metacognitive sensitivity, i.e. the precision with which subjects can monitor their own uncertainty.

Further details on the procedures for these two sensorimotor tasks, and the subsequent computation of measures of interest, can be found in their respective sections below.

3.2.3 Measuring declarative beliefs with the sense of agency scale

Subjects completed the Sense of Agency Scale (SoAS). Since this scale was introduced and first validated by Tapal et al. (2017), it has become widely used in the literature. Likert scale responses to each of the 13 items on the scale are multiplied by prescribed factor loadings

to obtain numerical values for sense of positive agency (SoPA) and sense of negative agency (SoNA), factors which explain separable components of variance in item responses/agency beliefs. While this decomposition may seem unintuitive – i.e., it is unclear why low SoNA should differ from high SoPA – this factor structure has been replicated at least three times in three different languages and populations (Bart et al., 2023; Hurault et al., 2020; Tapal et al., 2017). However, such a dissociation may make sense in light of the view that sense of agency is influenced by both by "positive" prospective cues – such as the ease with which one makes decisions (Sidarus and Haggard, 2016) – as well as "negative" cues such as unexpected sensory feedback (Synofzik et al., 2008).

The test–retest correlations reported by Tapal et al. (2017), measured two months apart, are $r = 0.78$ for SoPA and $r = 0.74$ for SoNA; these correlations may be useful to consider as a point estimate of the potentially explainable variance in scores when considering the effect size estimates we report here. Moreover, this high test–retest reliability lends itself to the interpretation that declarative agency beliefs captured by the scale are sustained over time. However, it is worth noting that while the SoAS is sometimes described in the literature as measuring "trait-level agency," more recent psychometric work has found that sense of agency can be further dissociated into two dimensions: situational – transcending an individual agency judgment but specific to a given context – and dispositional – context independent beliefs about one's own agency – both of which may contribute to subjects' scores on the SoAS (Di Plinio et al., 2024).

Subjects' data were excluded from this task if they only replied with 1's and 7's on the Likert scale or gave the same response to every question, indicating lack of honest effort in responding. 21 subjects were excluded on this basis, resulting in $n = 174$.

3.2.4 Intentional binding task

We use a conventional Libet clock paradigm (Haggard et al., 2002) for measuring the intentional binding. In this task, the subject sees a moving clock hand on each trial, and they are asked at the end of each trial to move the clock hand back to where it was when a "critical event" occurred; this event varies by condition. In the "baseline key" condition, the subject is asked to press a button on their keyboard during the trial (at their leisure but earlier than 8 s into the trial and but after the first rotation of the clock hand has occurred, or the trial restarts); the critical event is the objective time of the keypress. In the "baseline tone" condition, subjects are played an audio tone (at a random time, uniformly distributed throughout the trial), which is the critical event. In the "operant" conditions, the subject presses a key, and their keypress is followed by a tone a constant 0.25 s later. The critical event is the keypress in "operant key" trials and the tone in "operant tone" trials. The measure of intentional binding for each subject was computed separately for the key – "action binding" – and for the tone – "outcome binding" – by subtracting the average misestimation of the event onset (in milliseconds relative to the true event onset) in the baseline condition from that in the operant condition. Thus, these measurements reflect the degree to which perception of the keypress and tone events are shifted toward each other in time when the former is perceived as causing the latter, or intentional binding. Each 2x2 condition (i.e. baseline-tone, operant-tone, etc.) had 40 trials preceded by 5 practice trials, and we used a preexisting jsPsych implementation of the Libet clock paradigm which had already been validated for online use (Galang et al., 2021).

It is common in paradigms that measure intentional binding using a Libet clock paradigm (as opposed to an interval estimation paradigm) to report the perceptual shift corresponding to the action/keypress and outcome/tone separately, as key and tone binding may vary independently and may contain complementary information (Render and Jansen, 2021). In our case, we focus on action and outcome binding simply because the study which validated

the JavaScript implementation of the Libet Clock paradigm we used reported both, and we prefer using measures with validated implementations for timing-sensitive experiments (Galang et al., 2021). For analyses of individual differences (see 3.2.7 Data Analysis below), the sign of the individual subject outcome binding effect was flipped, such that more positive values always mean stronger binding – just as for action binding.

Subjects' data were excluded from this task if their over- or under- estimation in any one condition was farther than 5 standard deviations from the mean estimation in order to ensure included subjects were not responding randomly. Only 5 subjects were excluded on this basis, indicating decent task compliance overall. The resulting sample size was $n = 190$.

3.2.5 Control detection task

We used Wang and colleagues' jsPsych implementation of the sensorimotor task they introduced and validated (Wang et al., 2020). As with other control detection/discrimination tasks in the literature (Wen and Haggard, 2018, 2020), the task of the subject is to determine which of two moving dots they are able to influence the trajectory of by moving their mouse, while the actual degree of control is low enough to make accurate discrimination challenging. In this way, the task measures perceptual sensitivity to control cues.

Specifically, two moving dots, following independent, pseudorandom trajectories, were presented within separate circles on the screen. The subject could move their cursor to influence the trajectory of one of the two dots (the "target" dot) but were not told which dot they were influencing. The percentage of the target dot's trajectory that the subject could influence ("percent control") was manipulated across trials. Subjects had 4 s to view/influence the dot stimuli during which they could move their mouse as much as they wished, followed by a 0.5 s blank screen before they were asked to identify which dot they thought they were influencing. Subsequently, they were asked to rate their confidence in their answer. A video recording of several trials of this task is available with the online version of Wang et al.'s

(2020) paper.

As described by Wang and colleagues, the task begins with 5 practice trials starting at 25% control (very easy). After the practice trials, the experiment proceeded in two interwoven adaptive staircase procedures by which percent control was adjusted 13 times over 100 trials per staircase, resulting in a total of 200 trials. 15% of those 200 trials were randomly inserted "catch" trials, in which percent control was always 25%. Within each staircase, difficulty increased (i.e. true control level decreased) each time subjects got an answer correct and decreased each time they were incorrect. "Reversals" occurred whenever the staircase procedure switched from increasing to decreasing and vice versa. As the goal of the staircase was to hone in on the percent control in which the subject could identify the target dot with 75% accuracy, the amount of each difficulty increase or decrease following correct or incorrect trials, respectively, were asymmetrically weighted and decreased throughout the task to ensure the procedure asymptoted at 75% accuracy as described by Wang et al. (2020). (Please refer to Wang and colleagues' (2020) paper for a full description of the staircase procedure.) The average percent control along the last five staircase reversals was taken as the "percent control threshold," which served as our metric for each subject's sensitivity to visual control cues during sensorimotor agency judgements. As pointed out by Wang et al. (2020), the distribution of percent control threshold measurements is highly skewed, so these values were log transformed so as to be closer to normally distributed ("log control threshold"), and the log control threshold was used for analysis.

Moreover, to quantify metacognitive sensitivity, we computed meta-d' by maximum likelihood estimation (Fleming and Lau, 2014), using the *metadpy* Python package. This common measure of metacognitive performance reflects how calibrated subjects' uncertainty judgements are – that is, are they actually wrong more often when they are more uncertain? – while controlling for objective task performance.

Following Wang et al. (2020), subjects' data were excluded from this task if they failed

over 40% of the easy catch trials, indicating they were responding effectively randomly. Moreover, we excluded subjects whose uncertainty ratings scores were significantly below chance at predicting incorrect trials with a significance threshold of $p < 0.05$ via a Mann-Whitney test; that is, reported confidence was inversely correlated with success. We interpreted such cases as a misunderstanding of task instructions, such as believing 1 was "most confident" rather than "least confident," whereas subjects with whose certainty ratings were moderately but not significantly below chance were assumed to represent legitimate variance in performance. Subjects were also removed if they always gave the same uncertainty rating, as this prevented the calculation of metacognitive sensitivity measures (and presumably implied lack of effort). One additional subject was removed because their data contained unexplained missing values. Based on these criteria, 13 subjects were excluded, resulting in a sample size of $n = 182$.

3.2.6 Other self-report measures

We asked subjects to complete two other brief surveys for the purpose of obtaining pilot effect size estimates for future studies. Thus, these scales were never analyzed and results are not reported here, but the raw data are available in our open dataset and may be of use to other researchers. These scales were the Tellegan Absorption Scale (Tellegen and Atkinson, 1981) and the Embodied Sense of Self Scale (Asai et al., 2016).

3.2.7 Data analysis

All analyses and visualizations were done using Python. Distributions of measurements from the two sensorimotor tasks were visualized using the *DABEST* and *ptitprince* packages (see Fig. 3.1). Confidence intervals were derived by bootstrap for the Cohen's d effect size of intentional binding effects for purposes of replication. Before subsequent analyses of individual differences, all variables were z-standardized to facilitate the application of

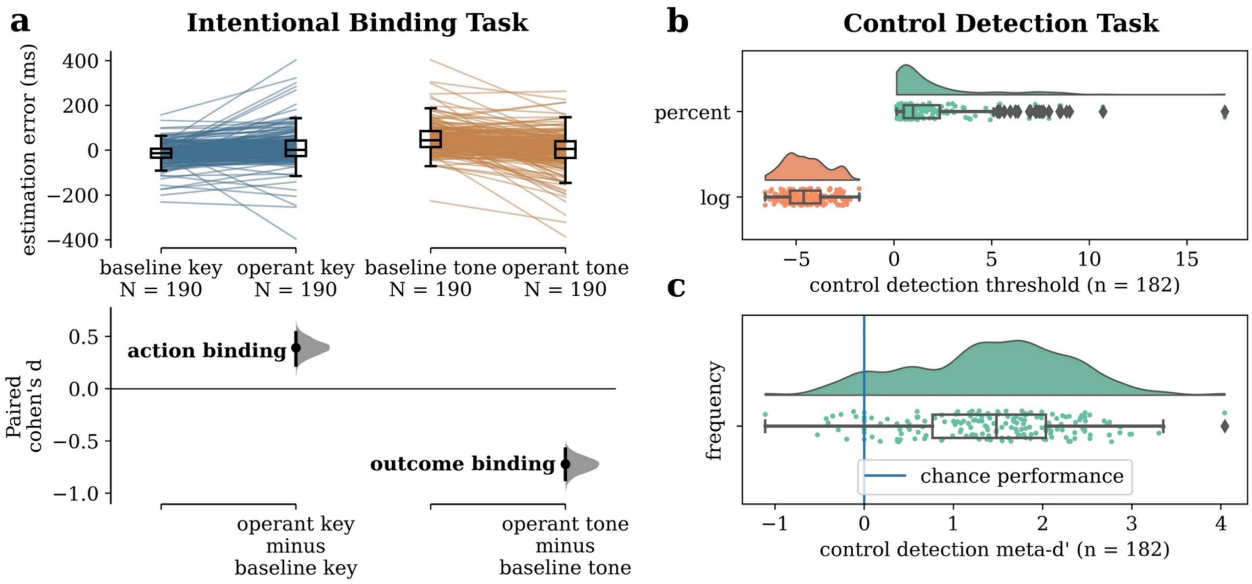

Figure 3.1: Distributions of sensorimotor, behavioral measures. (A) Each subjects' mean estimates of the timing of keypress and tone events relative to the true event times in the intentional binding task in each condition are shown on top, with bootstrapped distributions and 95% confidence intervals of the Cohen's d effect size for group-level intentional binding effects on bottom. (B and C) Raincloud plots of control detection threshold (in both percent and log scale), measuring sensitivity to control cues, and meta-d', measuring metacognitive ability. Box component of raincloud plots shows the median and quantiles, while whiskers show the extent of the distribution excluding extreme points (for visualization only). (B) Extreme points (those that fall more than 1.5 the interquartile range from the closest quartile are marked with diamonds; those same points are no longer extreme once log scaled.

conservative Bayesian priors (see Subection 3.2.8 below).

In our main analysis, we estimated correlations between each sensorimotor measure (i.e. action binding, outcome binding, log control detection threshold, and meta-d') and each measure of declarative beliefs (SoPA and SoNA). To test whether sensorimotor measures of sense of agency predict declarative beliefs about agency, we correlated each of the sensorimotor measures with both SoPA and with SoNA. All such correlations are estimated and reported, regardless of outcome, to avoid selection bias. We report full Bayesian posteriors for all correlations, which are used for inference (see Subsection 3.2.8 below).

Under the assumptions of signal detection theory, measures of first-order and of metacognitive sensitivity are mathematically guaranteed to be correlated (Galvin et al., 2003); in-

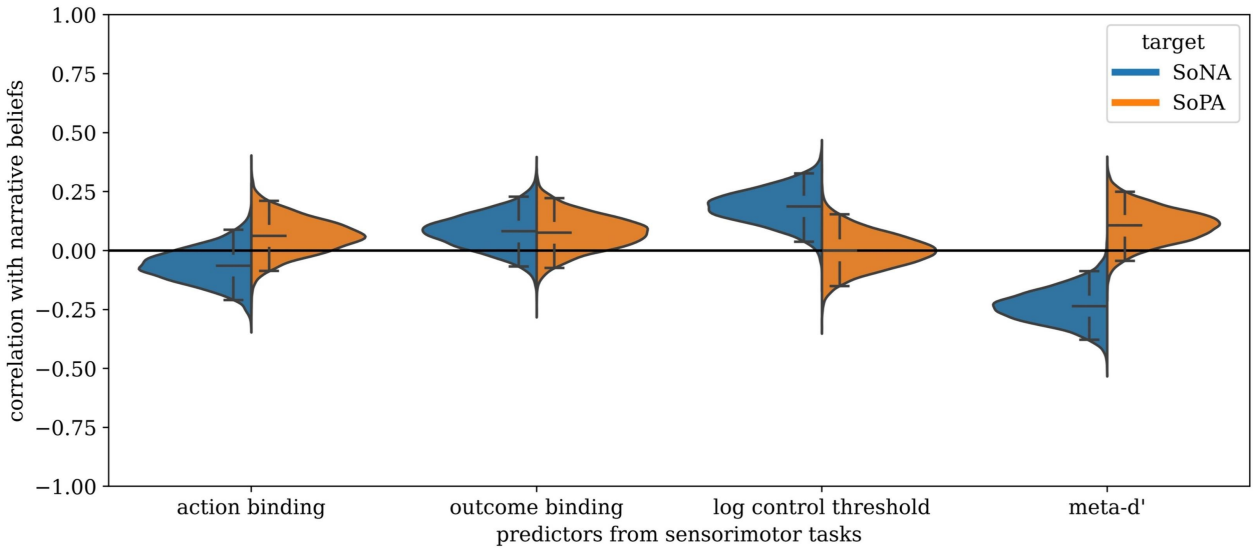

Figure 3.2: Posterior distributions of correlations between sensorimotor, behavioral measures and agency beliefs. Behavioral measures are as in Fig. 3.1. Agency beliefs, measured by the Sense of Agency Scale, are subdivided into sense of positive agency (SoPA) and negative agency (SoNA). Whiskers overlaid atop the violin plots extend to the 2.5% and 97.5% quantiles of the posterior distributions, representing 95% credible intervals.

deed, we did observe a correlation between log control threshold and meta-d' measures in the control detection task (see Fig. 3.3). Since both performance measures ended up being correlated with SoNA, we used mediation analysis to statistically control for this built-in correlation and ascertain which aspect of performance most directly accounted for their shared correlation with SoNA (see Fig. 3.3). Posterior distributions for total, direct, and indirect effects were estimated with two linear mediation models – one in which control detection threshold is the mediator, and another in which meta-d' is instead used as the mediator – and shown in Fig. 3.4.

3.2.8 Bayesian inference and multiple comparisons

The aim of a Bayesian data analysis is usually to estimate the probability of some unobserved parameters given the data, using Bayes' rule. The 95% "highest density interval" (HDI) or credible interval is often reported for each parameter, which is interpreted as reflecting a

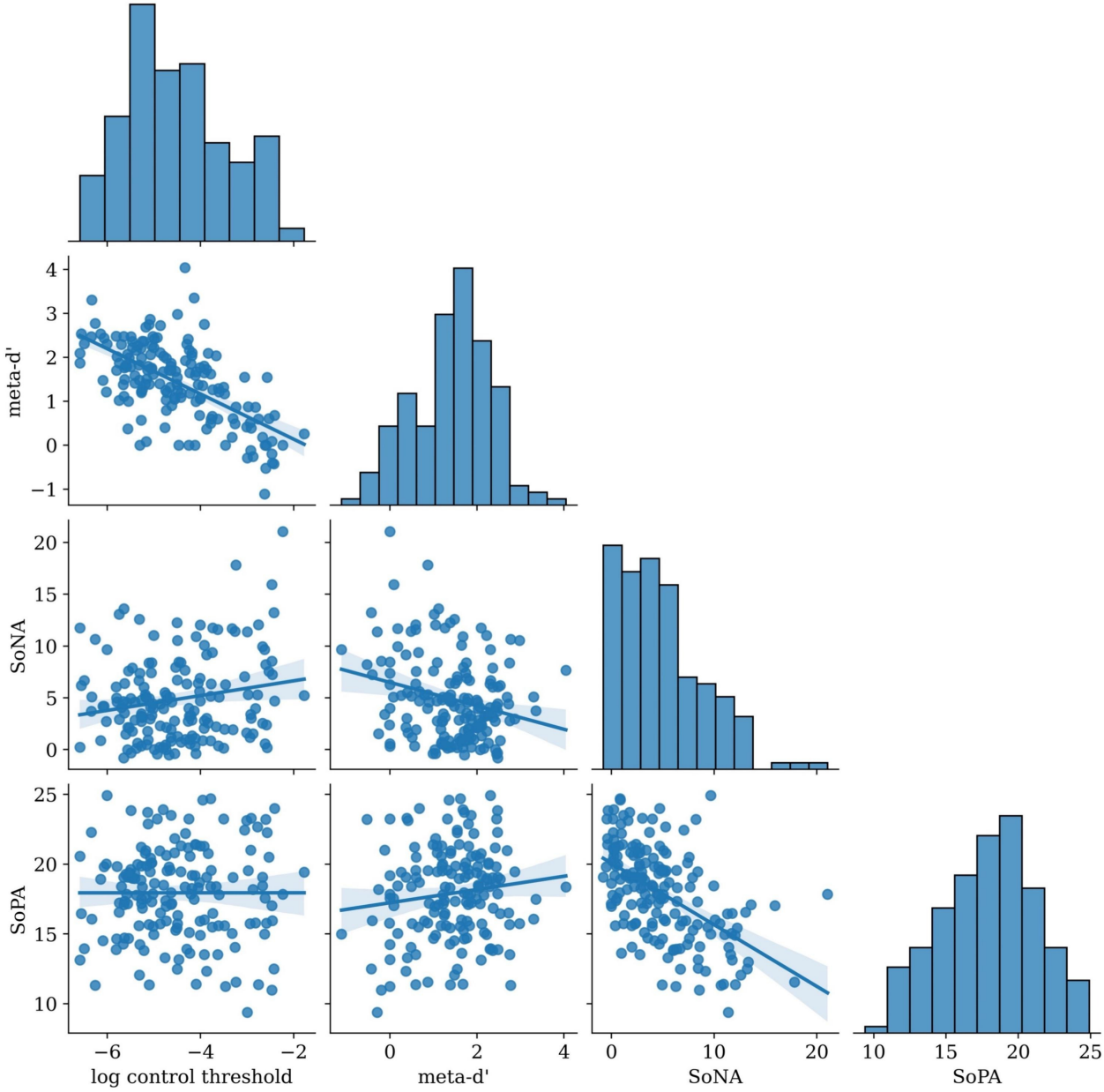

Figure 3.3: Joint distributions of control cue sensitivity, metacognitive performance, and sense of negative agency. Histograms for each variable are shown on the diagonal, raw data with best-fit linear regression lines and 95% confidence bands are shown off the diagonal. Data are shown in their original scale for visualization only.

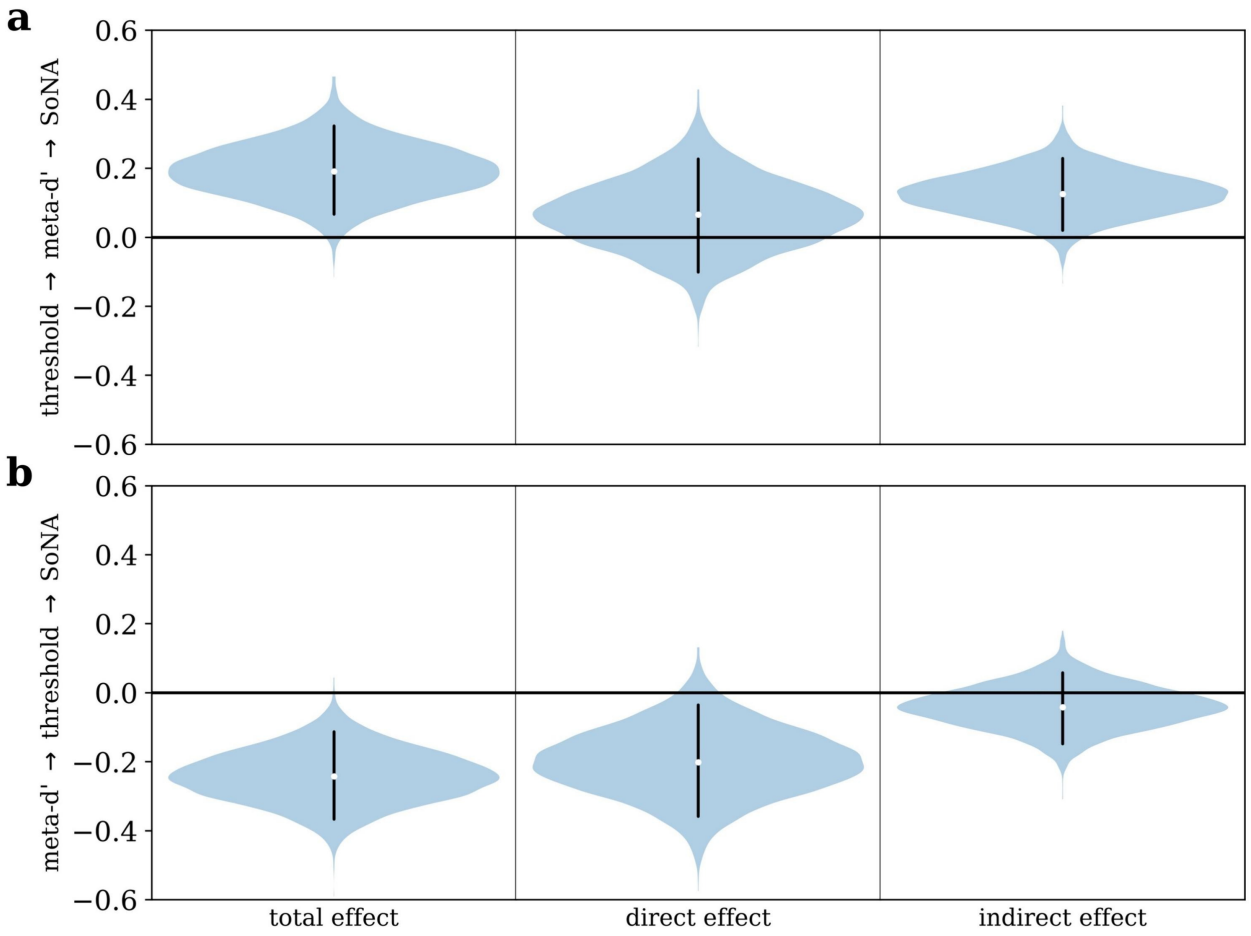

Figure 3.4: Posterior distributions for mediations analyses. (A) Posterior distributions and 95% HDIs for estimated effects of control detection threshold on negative agency beliefs (SoNA) with metacognition (meta-d') as a mediator. We find evidence for an indirect effect mediated by metacognition, but not of a direct effect. (B) Estimated effects of metacognition on SoNA with control detection threshold as a mediator. We find evidence for a direct effect of metacognition on negative agency beliefs, but not of an indirect effect. All variables (log control threshold, meta-d', and SoNA) were standardized before mediation analysis, so the regression coefficient estimates are on roughly the same scale as the correlations visualized in Fig. 3.2.

95% posterior belief that the parameter lies within that interval. (This is in contrast to a frequentist confidence interval, which is defined by a long term coverage rate over many repetitions of the whole study procedure – i.e. "If I compute such intervals on a very large number of repeated samples, 95% of them will cover the true parameter" – rather than probability that any given CI encompasses the true parameter.) It is typical, then, to interpret the data as showing evidence for a non-zero effect if the 95% HDI does not cover zero (or whatever the "null" effect size would be).

Since the denominator P(data) of Bayes' rule is normally intractable to compute analytically, but is conveniently known to be constant, modern Bayesian inference leverages the relative information in the numerator P(parameter|data)P(parameter) to draw samples from the posterior distribution, and the 95% HDI is approximated with an interval that contains 95% of these samples. Thus, for each parameter/model we estimate, we draw 10,000 posterior samples across 2 sampling chains using the no-U-turn sampler in Python's *PyMC* package.

If one uses "flat," uninformative priors P(parameter) in a Bayesian analysis, the HDI will end up being equivalent to a frequentist confidence interval, since P(data|param)P(param) will be proportional to P(data|parameter), the ordinary likelihood function. However, this is actually an undesirable effect, since many of the advantages of the Bayesian approach will be lost. Instead, if one sets "weakly informative" priors which rule out unreasonably large effect sizes a priori but are symmetric about zero to avoid biasing the analysis to show a non-zero effect, then effect size estimation is usually improved while heavily attenuating familywise error rate inflation due to multiple comparisons. For instance, if one z-normalizes their data before analysis (as we do here) and uses Normal(0, 1) (i.e. unit variance) priors for the linear relationship between two variables, then the familywise error rate inflation as the number of comparisons increases is empirically negligible, and parameter estimates are remain quite well calibrated (Gelman and Tuerlinckx, 2000). Intuitively, where the frequentist approach

Predictor	Target	Mean	Lower HDI	Upper HDI	Prob. Neg.	Prob. Pos.	p-val. (raw)	p-val. (FDR)
action binding	SoPA	0.063	-0.086	0.211	0.207	0.793	0.407	0.407
action binding	SoNA	-0.063	-0.207	0.090	0.798	0.202	0.400	0.407
outcome binding	SoPA	0.075	-0.070	0.225	0.157	0.843	0.315	0.394
outcome binding	SoNA	0.081	-0.069	0.225	0.146	0.854	0.280	0.394
log control thres.	SoPA	0.001	-0.154	0.150	0.498	0.502	0.997	0.831
log control thres.	SoNA	0.186	0.043	0.337	0.008	**0.992**	0.014	0.045
meta-d'	SoPA	0.106	-0.045	0.332	0.081	0.919	0.164	0.338
meta-d'	SoNA	-0.235	-0.378	-0.087	**0.999**	0.001	0.002	0.009

Table 3.1: Posterior summary statistics for correlations between behavioral measures and agency beliefs. Summary statistics include posterior mean (expected value), lower edge of 95% highest density interval (HDI), upper edge of 95% HDI, posterior probability effect size is negative, and probability effect size is positive. We also include raw and FDR-corrected p-values for the corresponding frequentist correlations.

to dealing with multiple comparisons is to effectively increase the width of all confidence intervals while keeping point estimates constant, the Bayesian approach is no shift the whole HDI toward zero. So for all of our analyses, we used a Normal(0, 1) prior for population means and regression coefficients, an Exponential(1) prior for variance/noise terms, and an LKJ(eta = 2) prior for correlations – all priors which will shrink effect estimates and their HDIs toward zero.

For readers who are uncomfortable with the Bayesian approach to multiple comparisons correction (or who simply wish the see the results of an analysis approach with which they are more familiar), we also provide raw and false-discovery-rate (FDR) corrected p-values (using a Benjamini-Hochberg FDR correction with $\alpha = 0.05$) for the frequentist correlation corresponding to each of our Bayesian correlations in our summary table, though they are not discussed in-text.

3.2.9 Data and code availability

All code for both the experiment (https://doi.org/10.5281/zenodo.8173285) and the analysis (https://doi.org/10.5281/zenodo.8173282) is permanently archived on *Zenodo*. Deidentified raw data is available on the *Open Science Framework* (https://osf.io/753c2/) and is organized roughly according to the Brain Imaging Data Structure specifications for behavioral data to facilitate easy navigation.

3.3 Results

We observe distributions of behavioral effects consistent with the prior work from which our sensorimotor tasks (intentional binding and control detection) were taken (Galang et al., 2021; Wang et al., 2020). We replicate the previously reported intentional binding standardized effect sizes for both key (d = 0.39, 95% CI [0.23, 0.54]) and tone (d = -0.72, 95% CI [-0.87, -0.58]), corresponding to absolute effect sizes of 27.85 ms (95% CI: [17.68, 39.25]) and -57.89 ms (95% CI: [-71.27, -46.57]), respectively. In comparison, the meta-analytic effect sizes for the intentional binding effect are d = 0.45 and d = -0.73 for action (key) and outcome (tone) binding, respectively (Tanaka et al., 2019). In other words, we obtained effect size estimates consistent with "gold standard" in-lab measurements. Additionally, the distributions of percent control threshold and metacognitive sensitivity we observe in the control detection task are similar in shape and in mean (percent control threshold = 2.07, 95% CI: [1.74, 2.50]) to those obtained by Wang and colleagues in their original validation of the task (Wang et al., 2020). Meta-d' measurements had mean 1.41 (95% CI: [1.27, 1.54]). Observed distributions of behavioral measurements of interest are visualized in Figure 3.1.

Bayesian posterior distributions for Pearson correlations between potential sensorimotor predictors and measures of declarative beliefs are shown in Fig. 3.2, with summary statistics in Table 3.1. We find evidence of a positive correlation between log control detection threshold and negative agency beliefs (r = 0.186, 95% HDI: [0.043, 0.337]). That is, those who are less sensitive to control cues report having less agency. Moreover, we find evidence of a negative correlation between metacognitive ability (meta-d') and SoNA (r = -0.235, 95% HDI: [-0.378, -0.087]). That is, those with better metacognitive ability for agency judgements report feeling less negative agency overall. We do not find sufficient evidence to draw a conclusion as to whether intentional binding magnitudes predict beliefs about agency, but we report 95% "highest density intervals" (HDIs, i.e. Bayesian credible intervals) in Table 3.1 which place upper bounds on how large such an effect could plausibly be based on our

data.

While we found that control sensitivity (log control detection threshold) and metacognitive performance (meta-d') both correlate with SoNA separately, visualizing the joint distributions of the three measurements as in Fig. 3.3 reveals a clear correlation between log control threshold and meta-d' of r = -0.622 (95% HDI [-0.708, -0.521]) between the two measures. This finding motivated a mediation analysis to determine whether this correlation between behavioral predictors confounded our estimate of their correlations with SoNA.

In a mediation model in which log control threshold is a predictor and meta-d' a potential mediator, we find evidence that the total effect of control sensitivity on negative agency beliefs (beta = 0.190, 95% HDI: [0.037, 0.340]) can be explained in part by an indirect effect mediated by metacognition (beta = 0.130, 95% HDI: [0.003, 0.250]), but we do not find sufficient evidence of any direct effect (beta = 0.066, 95% HDI: [-0.120, 0.260]). In contrast, when we use metacognition as a predictor and control threshold as a mediator, we find that the total effect of metacognitive ability on negative agency beliefs (beta = -0.240, 95% HDI: [-0.390, -0.090]) can be explained by a direct effect (beta = -0.200, 95% HDI: [-0.390, -0.006]), and we do not find evidence of an indirect effect (beta = -0.042, 95% HDI: [-0.160, 0.084]). Taken together, we can conclude with high probability that metacognition mediates the relationship between control detection threshold and negative agency beliefs; conversely, there is an effect of metacognitive ability on negative agency beliefs that is not mediated by the control detection threshold (see Fig. 3.4).

While we model control sensitivity as affecting agency beliefs, please note our analysis does not rule out the possibility that causality may flow in the reverse direction as well, which would generate an identical partial correlation matrix (Fiedler et al., 2011). Our mediation analysis simply constrains the possibility space of causal structures relating control sensitivity and agency beliefs (with whatever directionality) to those that are mediated by metacognition.

3.4 Discussion

Agency judgements (or self-vs-other judgements in general) do not only occur at the sensorimotor level, nor is there a clear boundary between experiences of individual agency judgments and declarative beliefs. Sense of agency (SoA) has been studied at many levels of abstraction (e.g. mental, social, etc.), and recent controversies have cast doubt on the notion that a common cognitive or neural substrate can account for agency judgements across all these scales (Charalampaki et al., 2022; David, 2012; Gallagher, 2012; Pacherie, 2007). Indeed, the neural predictors of agency judgements appear to meaningfully differ even between different types of sensorimotor judgements, such as those concerning muscle movements (Veillette et al., 2023b) and downstream outcomes (Timm et al., 2016). If asked, however, most people might say that the "I" to which they attribute actions and consequences does not differ across these domains; this unity of self-as-agent in experience seems to contradict the heterogenous cognitive and neural mechanisms that account for SoA across levels of abstraction. While our results are indeed consistent with the view that subjects' sensitivity to their own control in the sensorimotor domain is not congruent with their declarative beliefs about agency, we do find evidence of a surprising relationship between the two.

While it would be intuitive to theorize that those who experience agency more frequently in their moment-to-moment agency judgements will report higher SoA when asked about their beliefs about their own agency – as a matter of statistical learning – what we find instead is more nuanced. While the intuitive correlation between (in)sensitivity to control cues (i.e. control detection threshold) and sense of negative agency (SoNA) does appear to exist, this effect was mediated primarily by metacognitive accuracy (i.e. meta-d', see 3.2 Methods) about such agency judgements – that is, the accuracy with which one monitors uncertainty about agency judgements. In other words, we do not find evidence of a direct relationship between sensitivity to control cues and declarative agency beliefs but of an indirect relationship mediated by metacognition. While sensitivity to control and metacognitive accuracy

are correlated in the present study, this need not be the case in all situations; evidence suggests agency judgements can be made without recruiting metacognitive resources (Constant et al., 2022). In such cases, however, metacognition may still play a role in determining how individual experiences of agency are integrated into a persistent self-concept.

We do not find convincing evidence for a correlation between intentional binding measurements and either measurement of SoA beliefs. An advantage of Bayesian analyses is that they sometimes allow one to interpret null findings (i.e. when the HDI is narrow around an effect size of zero); however, since some of our 95% HDIs comfortably contain correlations as large as 0.2 (see Table 3.1), we also cannot rule out substantial, non-zero correlations in this case. Interestingly, previous work has suggested that the magnitude of intentional binding predicts free will beliefs (Aarts and van den Bos, 2011) and, conversely, that free will beliefs affect motor preparatory neural activity (Rigoni et al., 2011). Given that free will beliefs are correlated with the intentional binding effect, one might expect that agency beliefs would be as well – as beliefs in one's own causal power intuitively seem to be a special case of the belief that people have causal power in general. SoA beliefs, which pertain specifically to one's own ability to exert control over the world, do however vary independently of free will beliefs (Tapal et al., 2017). One possibility which may explain this difference is that free will beliefs are more related to phenomenal correlates of agency (i.e. how it feels to produce actions/outcomes), as reflected in intentional binding, rather than whether or not actions are believed to be self-caused. This distinction could be a fruitful subject for future study.

Moreover, we did not find compelling evidence that any sensorimotor metric predicted positive SoA (SoPA), only SoNA. This finding (or lack thereof) makes sense in light of existing theory, as interruptions of normal sensorimotor control become salient intrusions in conscious experience, but the routine flow from action to outcome naturally falls into the background (Synofzik et al., 2008). However, another possibility is that SoPA items on the SoAS showed worst internal reliability (Cronbach's alpha = 0.651, 95% CI: [0.562, 0.727])

than did SoNA (Cronbach's alpha = 0.838, 95% CI: [0.797, 0.872]); in other words, SoPA scores could have just been noisier, allowing a true correlation to evade detection.

It is important to note some limitations on the inferences we can draw from the present data. Obviously, we did not measure all possible behavioral indices of agency experience at either the sensorimotor or the narrative level; indeed, no single study can. Consequently, we cannot rule out a direct effect of control sensitivity or some other index that would affect the frequency of positive agency judgements on SoA beliefs. Not all such relationships, if they exist, are necessarily mediated by metacognition. Of particular note, the sensorimotor tasks used here primarily reflect the action-to-outcome part of the intention-action-outcome chain, while prospective cues that are known to influence SoA seem to arise from the intention/selection process itself (Haggard, 2017). Moreover, our correlation estimates fall far below the test–retest correlation of the Sense of Agency scale (Tapal et al., 2017), suggesting that there is still much meaningful variance in agency beliefs left to be explained – in all likelihood by factors that are not to be found at the sensorimotor level. Moreover, while the mediation analysis was necessary to tease apart the contributions of first order and of metacognitive sensitivity to predicting agency beliefs, as these two variables were mutually confounding, this analysis was not pre-planned; it would be beneficial for future research to replicate these findings in a new sample.

Further, the extent to which the observed correlations are explained by a causal effect of sensorimotor experience on beliefs, rather than of beliefs on sensorimotor experience, remains unclear, as mediation analyses (or any correlational analysis) cannot distinguish between those two causal models, which would give rise to identical partial correlation matrices (Fiedler et al., 2011). In other words, while we can say with some certainty that the observed relationship between control sensitivity and agency beliefs is through metacognition, these data alone cannot tell us whether moment-to-moment agency is driving beliefs or if beliefs influence moment-to-moment agency. In the former direction, a likely interpretation is that

only sensory evidence of control to which one has metacognitive access – rather than all the sensory evidence that affects the first order decision – is ultimately incorporated into agency beliefs; this is consistent with the interpretation of the meta-d' measure under signal detection theory as described in 3.2 Methods (Fleming and Lau, 2014). Conversely, we have more trouble imagining a mechanism by which agency beliefs could impinge upon first order agency judgments in a manner that is mediated by metacognitive sensitivity, as metacognitive reflection/introspection in principle occurs after those first order judgements. Thus, the present authors favor an interpretation in which sensorimotor-level experiences are integrated into declarative beliefs contingent upon metacognitive access, but we want to make clear that is a theoretic interpretation of the data rather than a fact implied by the data themselves.

However, our results clearly show that a (surprisingly) substantial portion of the individual differences in self-agency beliefs are concretely related to the perception of volitional action in a sensorimotor setting, and that the observed relationship is mediated by metacognition. We interpret our findings as pointing toward a model of SoA in which moment-to-moment experiences of agency are aggregated into beliefs contingent upon having metacognitive access to the evidence of control that resulted in those experiences (though we cannot rule out, from our data, the possibility that beliefs instead or also impact agency judgments through metacognition). Recent evidence has suggested that first order agency judgements may not always recruit metacognitive resources, leaving the role of metacognition in the experience of agency unclear (Constant et al., 2022). The present study suggests it may serve a critical function in incorporating experiences of agency into narrative beliefs.

CHAPTER 4

DISCUSSION

Taken together, the preceding chapters provide substantive support for the claim that the conscious sense of agency, even including declarative beliefs about oneself as an agent, is influenced by the neural and cognitive processes that coordinate bodily movement. In this way, the sense of agency provides a consciously accessible index into these sensorimotor processes that otherwise operate largely outside of awareness. Such a conscious window into the efficacy of ongoing motor processes would afford an organism the ability to update consciously controlled behavioral strategies. Consistent with this functionalist view, humans demonstrate explicit choice behavior consistent with an intrinsic preference for agency, selecting higher agency environments even at the cost of monetary reward (Norton and Liljeholm, 2020). In the long run, this heuristic may nonetheless be reward maximizing if an organism is uncertain about its own future preferences, in which case preserving the ability to alter outcomes in the future is advantageous (Liljeholm, 2022). In fact, self-agency has been proposed as an intrinsic reward function for reinforcement learning, with some empirical benefits demonstrated for (robotic) motor learning (Leibfried et al., 2019).

That said, findings presented in the preceding chapters do not strongly suggest which specific sorts of motor representations directly influence the conscious sense of agency, merely that some do. While I do present evidence for a specific neural circuit by which actions and outcomes may be paired in Chapter 1, I did not present evidence that this specific process, or any other specific process, actually impinges upon conscious awareness. The results in Chapter 2 clearly suggest that sensorimotor neural processes are linked to the conscious sense of agency, as agency judgments can be predicted from related neural activity on a trial-by-trial basis. However, results do not point to a specific component of sensorimotor processing. As discussed in detail in Chapter 1, current computational models of motor control minimally include a "forward dynamics model" by which an organism predicts future

sensory (or perhaps bodily) states that would result from some action, an "inverse dynamics model" or control policy which selects actions to achieve a desired state, and a prediction error indexing the difference between the forward model's predictions and actual sensory feedback (Wolpert and Ghahramani, 2000). Multiple paired forward and inverse models would likely be necessary, in this framework, for different sensorimotor environments, across which physical dynamics differ (Wolpert and Kawato, 1998; Heald et al., 2021). (For example, given the same muscle activations, resulting movements will be quite different for somebody standing on dry land from the same person swimming in water.) Given the extremely high degrees of freedom of the musculoskeletal system, each of these bidirectional environment models would in principle need to be very complex, seemingly implying a vast array of internal representations. From this perspective, it is natural to ask which categories of these numerous internal representations influence the sense of agency, which could act as a high-level index of the efficacy of one's models in the current environment.

In more recent work, I have begun developing a general framework for generating specific, testable predictions about brain activity measures from theories that relate categories of representations found in models of motor control to elements of subjective motor awareness. In essence, the idea is to embed an algorithmic model of motor control into a musculoskeletal body within in a physics simulation. The model learns from sensorimotor experience in the physics simulation until it can perform a motor task, and a human participant performs the same task whilst brain activity is measured (e.g. using fMRI). Where the algorithm's learned representations linearly predict out-of-sample brain activity better than nonlinear transformations of the task state (e.g. body position, target position, etc.) fit directly to some training data, this is interpreted as a sort of localizer for the hypothesized motor process approximated by the algorithm (see Fig. 4.1). In a separate session, the same participant performs a task designed to measure or manipulate some aspect of motor awareness, as the task from Chapter 1 manipulates sense of agency. If self-reports or implicit measures of a

subjective experience of interest would be predicted from the areas/times identified by the localizer, this would be interpreted as evidence of an association between the hypothesized motor control process modeled by the algorithm and the subjective experience of interest. Algorithms can be designed to satisfy particular computational specifications with respect to their role in the artificial agent's architecture, such that they can be made to approximate specific components of computational-level descriptions of motor control (e.g. forward model, inverse model, etc.). Thus, the same dataset may be used to test between multiple computational/algorithmic models of motor control, based on how well they predict brain activity on one hand, and explain aspects of conscious motor awareness on the other.

Our first (to some degree intentionally naïve) attempt at implementing the above approach in practice has already yielded interesting results suggesting a non-visual role of early visual cortex in inverse dynamics computation for neuromuscular control of the hand (Veillette et al., 2025a), but that positive finding is out of scope of this dissertation. More pertinent to the current Discussion is our negative findings, where our localizer failed to localize representations that it in principle should have. Specifically, the output of our approximate "inverse dynamics model," which was trained to control a musculoskeletal hand in simulation using deep reinforcement learning, is a vector of muscle activations; since the corticospinal output of motor cortex strongly resembles downstream electromyographic activity (Marshall et al., 2022), one would intuitively expect the output representations of the algorithmic model to predict neural activity in motor cortex. While the model representations learned in simulation do, indeed, predict fMRI-measured brain activity in motor cortex, the control model – which learns simple mappings from task state to brain activity directly – performs much better (see Fig. 4.1c). Taking our localizer at face value, then, leads to a biologically implausible conclusion. However, it is not entirely unexpected that this approach failed.

While an important function of motor cortex is to drive muscle activity downstream

of the corticospinal tract, the activity of only a very small fraction of variation in motor cortex population activity actually resembles muscle activity. Instead, much of the variance in motor cortex activity is contained within an "output null" subspace, in which variation in motor cortex activity does not seem to correspond to measurable (or at least as-of-yet measured) changes in corticospinal output; this is understood to prevent motor preparatory activity from causing unintentional movements (Churchland and Shenoy, 2024). To initiate a movement, motor cortex activity must exit the output null subspace to being producing muscle activity; since population activity at a given time is not independent of the activity immediately preceding it, motor cortex activity must evolve over time along a trajectory from the output null to an "execution subspace." As a result, the single largest component of variation in motor cortex population activity – and all one can see with fMRI, as used for our localizer, is the population aggregate – is movement onset (Kaufman et al., 2016). Our control model, which learns mappings between brain activity and task state (which includes movement) can easily learn this simple feature, to which our inverse-dynamics-approximating model did not have access. Even after leaving the null space, trajectories of motor cortex activity do not resemble muscle movement trajectories but rather proceed in a quasi-rotational trajectory, which is thought to minimize movement error by ensuring a perturbation to neural activity does not knock population activity onto a "tangled" trajectory for a different movement (Russo et al., 2020).

The failure of our localization approach, then, was the result of a category error of a kind which was well articulated by David Marr (Marr, 2010). The notion of an inverse dynamics model (or a forward dynamics model, or a prediction error, et cetera) is a computational-level description of motor control. The computational specifications given by such a description may be satisfied by a number of possible algorithms. Algorithms, in turn, could potentially be implemented by a number of possible physical implementations. Problems arise when one confuses these levels of description. To construct our localizer, we selected a computational

specification: an inverse dynamics model, which takes a current and a target bodily state as input, and outputs a set of muscle activations. We also exercised flexibility in selecting an algorithm that satisfies that computational constraint: a deep neural network, trained using reinforcement learning. This algorithm treats inverse dynamic computation as a series of transformations between representations, proceeding consecutively from input to output. We then essentially correlated these representations, at each point of time during the task, with brain activity over time. However, as pointed out above, neural activity during movement does not passively represent muscle activity but generates it though a time-evolving process. Our approach, though it yielded promising preliminary results (Veillette et al., 2025a), fundamentally erred by conflating a representation from an algorithm-level description with a physical, implementation-level phenomena. This is not to say algorithm-level descriptions cannot be meaningful models of neural computation; however, neural activity should not be interpreted as directly reflecting representations but as reflecting the network-level dynamics that drive representations, as motor cortex drives muscle activity.

This category error certainly impedes progress in my particular aim of linking specific motor representations with the subjective experience of agency, but I would argue it is more generally a central obstacle to the progress of cognitive neuroscience as a whole. Recent years have seen a rapid rise of a "computational cognitive neuroscience" subfield; a stated aim in the paper that is often credited with delineating the computational cognitive neuroscience research program is, presciently, to integrate across Marr's levels of analysis by marrying computational models of behavior with cognitive neuroimaging (Kriegeskorte and Douglas, 2018). However, the analytic tools which have become the workhorses of computational cognitive neuroscience are, I argue, fundamentally incompatible with this end goal. Multi-voxel/variate pattern analysis (Kriegeskorte et al., 2006), multivariate predictive models (Kragel et al., 2018), voxelwise encoding models (Huth et al., 2016), and representational similarity analysis (Kriegeskorte et al., 2008) are routinely interpreted through the core as-

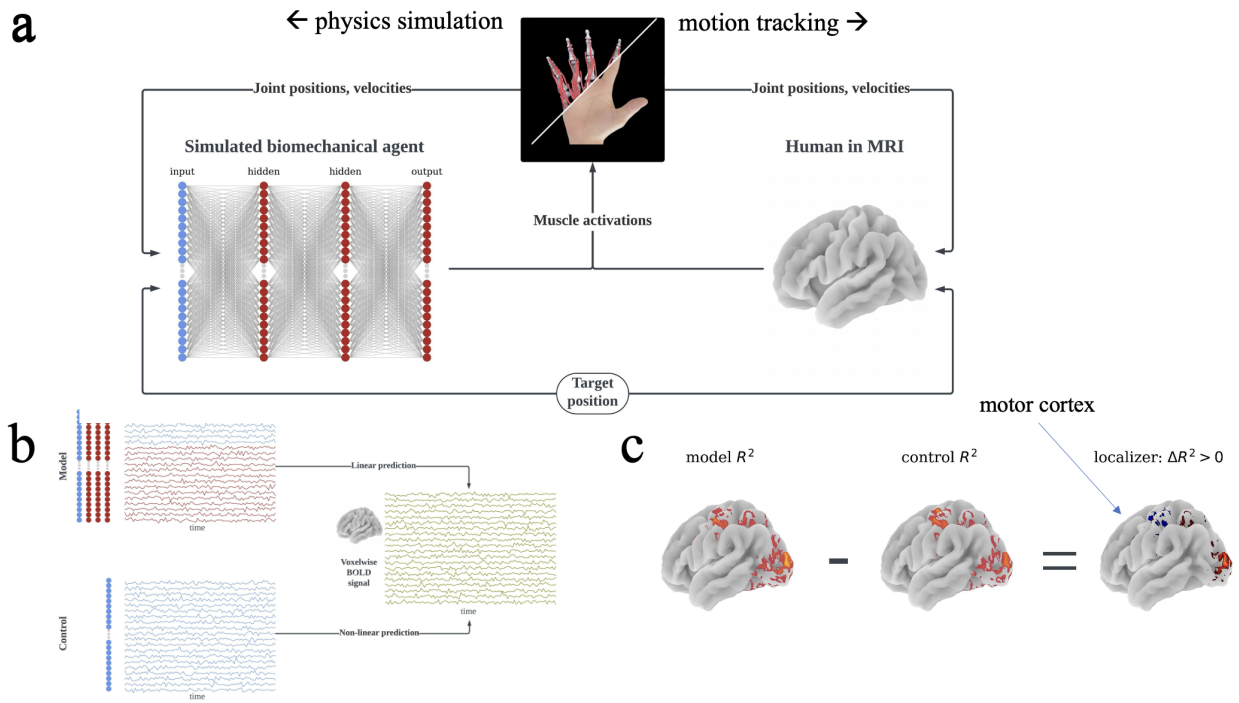

Figure 4.1: A model-based localizer for motor control representations. (A) An algorithmic agent model, such as a neural network, can learn to control a biomechanical body from sensorimotor experience in a physics simulation. (B) The representations the agent model learned in the physics simulation can be used as features in an encoding model to predict measurements of brain activity, such as the blood oxygen level dependent (BOLD) signal measured in fMRI, while a human participant performs the same task. However, it is important to compare the performance of the agent model's representations to that of a control encoding model which can capture simple nonlinear transformations of the same motor task. This ensures that high predictive performance is specific to the agent model. (C) We have previously attempted this approach in Veillette et al. (2025a), from which this figure was adapted, but the agent model failed to outperform the simple data-driven control model at predicting motor cortex activity. Taken at face value as a localizer, this result would lead to the biologically implausible conclusion that muscle activations (the output layer of the agent model) are not related to motor cortex activity. The result then, like many out-of-the-box applications of current computational cognitive neuroscience methods, should not be taken at face value.

sumption that measured neural activity reflects representations in a (usually linear) manner. In this view, representations are "encoded" as patterns of neural activity and can thus be "decoded" by the researcher, revealing the content thereof (Mathis et al., 2024). I myself often adopt the encoding/decoding terminology, as in Chapters 1-2, to describe particular data analysis approaches; however, a method of analysis should be distinguished from a theoretical claim. Indeed, the notion that the decodable contents of neural signals are substitutable for that which the brain represents has been articulated as a theoretical framework where neural representations are characterized by "representational geometries" (Kriegeskorte and Kievit, 2013; Chung and Abbott, 2021). These geometries can be quantified by "representational distance matrices" where the discriminability between neural patterns that coincide with various different stimuli, states, or behaviors is treated as a proxy for the difference between their representations. The failure of our localizer (Veillette et al., 2025a) illustrates the fallacious nature of this framework; a scientist taking this view, without knowing anything else about motor cortex, would arrive at the false conclusion that the primary function of motor cortex is to represent movement onsets (Kaufman et al., 2016).

Of course, analytic approaches that decode/predict stimuli and behavior from brain activity, or vice versa, remain useful when considered for what they are. And, at times, it may be true that the discriminability between neural patterns neatly corresponds to a meaningful, algorithm-level representation. Had Hubel and Wiesel (1959) adopted modern representational geometry assumptions when characterizing receptive fields, already essentially encoding models, of single cells in striate cortex, they would likely have drawn the same conclusions as they did more than half a century ago. More generally, I suspect that when studying passive responses to a stimulus, the representational geometry assumptions will often be quite benign. However, the activities of areas such as motor cortex are dominated by their rich, internal dynamics precisely because such subsystems need to generate behaviorally-relevant output, not just passively encode information. As a consequence, their

activities need geometrically reflect neither stimuli, behavior, nor any intermediate representation but instead the network-level strategy for contributing to robust, whole-organism behaviors (Saxena et al., 2022). Representational geometry axioms are not just violated in motor cortex, however. Much recent work in theoretical neuroscience has highlighted the robustness properties afforded by the topological properties of "dynamical attractors," manifolds to which neural activity are effectively constrained by their internal dynamics (Langdon et al., 2023). For example, grid cells in the entorhinal cortex have periodic receptive fields for spatial position; by coding for positions with the phases of individual cells' characteristic spatial frequencies (Gardner et al., 2022). If one assumes neural representations of position would be reflected as Euclidean distances between positions, as in the framework of representational geometry, they would hardly conclude that grid cells represented spatial positions; in fact, far-away locations may have similar or even identical corresponding population activity (see Fig. 4.2). However, the derivative of each cell's firing rate with respect to condition is not determined by its current firing rate, but by the current phase of its receptive field. Thus, the state-dependent dynamics of the cell population are not defined in a Euclidean space spanned by the firing rates, but rather in a toroidal space spanned by the phases (Gardner et al., 2022). In the native geometry of the torus attractor, the organism moving in a single spatial direction results in a one-dimensional trajectory of population activity. Euclidean space ignores the continuity constraint that, to transition from one phase of its receptive field to another, each cell must pass through all the intervening phases. Consequently, trajectories along a spatial dimension in Euclidean space are tangled such that nearby population states correspond to far away locations. This peculiar organization serves a purpose; by coding (in the representational geometry sense) distal positions adjacently, a noise perturbation could only move the population to a pattern corresponding to an implausible location, so if downstream place cells enforce a temporal continuity constraint (as captured by the toroidal geometry), the behaviorally-relevant representation remains unaffected (Fiete et al., 2008;

Burak and Fiete, 2009; Sreenivasan and Fiete, 2011). Thus, while "code" or "encoding" in the sense of the representational geometry framework does have a useful meaning here, the encoding is clearly distinct from the meaningful content of the representation, which is only well-defined in the context of the dynamics with which the momentary code may evolve over time.

With this lesson in mind, it is worth returning to a challenge posed in this dissertation's Introduction. The comparator model, on which most modern accounts of sense of agency are still based, posits that agency is high when the difference between the predicted and actual sensory consequences of movement is low (Feinberg, 1978; Frith, 1987). While this model was originally proposed to explain "thought insertion" in schizophrenia, in which one's thoughts may be perceived as originating from an external agent, it has failed to do so due to a lack of clear distinction between an internal prediction of a thought and an actual thought (Synofzik et al., 2008; Frith, 2012). I suggested, in the Introduction, that this limitation might be overcome by considering neurocognitive components of motor control other than just prediction error as potential contributors to the conscious sense of agency. I then suggested, in Chapter 1, a neural circuit hypothesized to match sensations with precipitating motor events based on temporal coherence between anticipatory neural representations of critical events in peripheral biomechanics and the onsets of resultant sensory feedback events, raising electrophysiological evidence from speech motor control and parallel observations from the birdsong model system (Amador et al., 2013). One might wonder how this mechanism would solve the aforementioned problem, since it merely replaces a prediction of content with a prediction of timing. If we assume neural patterns at a given moment constitute a code, and a code is equivalent to a representation, then indeed the problem remains. However, this tension could be resolved when considering the dynamics of the integrated sensorimotor system, which constrain how neural activity can evolve over time during movement production.

In one model of birdsong production – where hierarchically organized premotor cortex

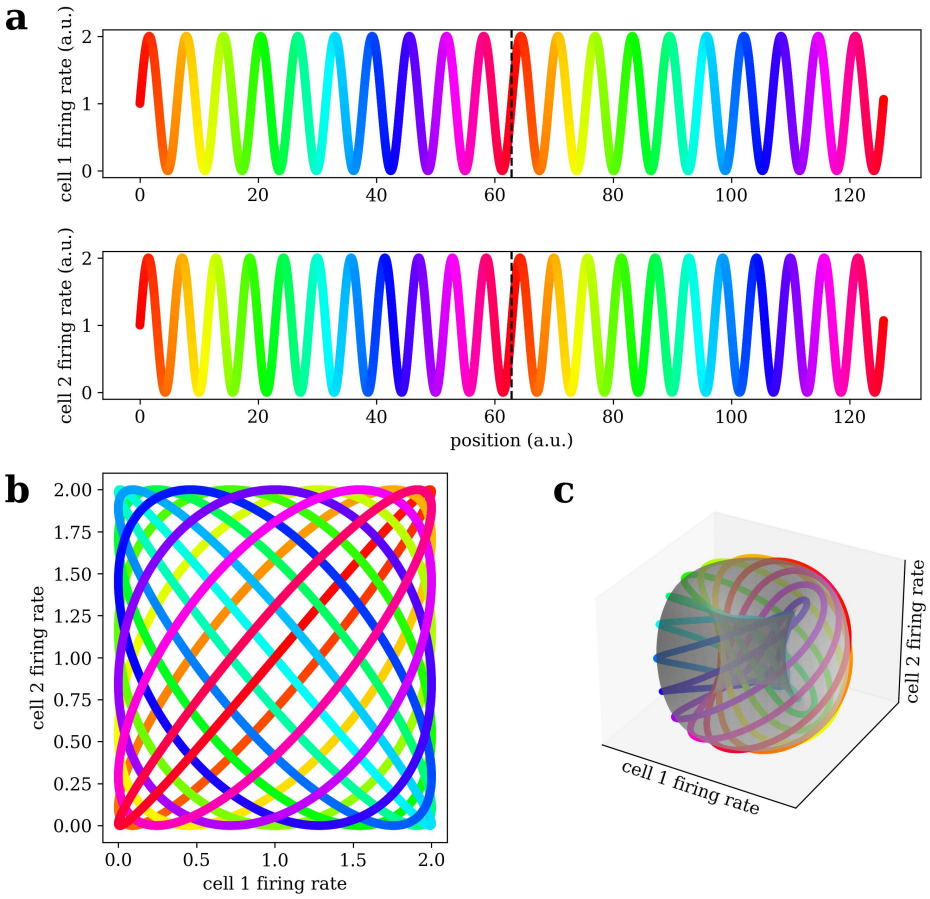

Figure 4.2: A simplified depiction of a grid cell population code. (A) The firing rate of a grid cell is a periodic function of spatial distance, here shown in just one dimension for simplicity. Two grid cells with receptive fields of slightly different spatial frequencies ω_1 and ω_2 can collectively distance, but only up to one period of their $|\omega_1 - \omega_2|$ beat frequency before their joint-phase code repeats, denoted with a dotted line. (B) If a "state," "multivariate pattern," or "population vector" in the Euclidean space spanned by the firing rates is taken as equivalent to the mental representation the grid cell population encodes, one would conclude that grid cells interpret very distant locations as nearby or sometimes even the same. (C) However, the derivative of each grid cell's firing rate with respect to spatial position is not a function of the cell's current *firing rate* but rather the *phase* of their receptive fields. Thus, the dynamics are not properly defined in the Euclidean space of activations but on the toroidal space defined by the combination of phases (the same way a single phase can be said to lie on a circle). On the torus attractor to which grid cell activity is confined by the internal dynamics of the entorhinal cortex (Gardner et al., 2022), an organism's one-dimensional path through space corresponds neatly to a untangled, one-dimensional neural population trajectory. The neural *representation*, then, is defined by the state-dependent dynamics of the population, not by the states themselves.

analog HVC, motor cortex analog RA, and respiratory-related brainstem area ER, as well as an unspecified "initiating area" which projects to both HVC and ER are all modelled as standard Wilson-Cowan equations (Wilson and Cowan, 1972) – simulated ER activity closely resembles empirical air sac pressure patterns when, and importantly only when, model parameters are set such that HVC, RA, and ER exhibit synchronous activation peaks (Dima et al., 2018; Alonso et al., 2015). Interestingly, this model predicts from first principles the empirical observation that neurons in adult zebra finch HVC burst synchronously to extrema in air sac pressure (Amador et al., 2013). A similar phenomenon to that finding and to Chapter 1's results in human speech production has been observed in juvenile zebra finches, where HVC activity appears to be synchronized to rather than precede syllable onsets (Okubo et al., 2015), suggesting that such synchrony is not dependent upon the adult-level vocomotor fluency. The HVC activity that initially drives the respiratory brainstem through RA, in this model, culminates in activity that is synchronous to actual respiration as a deterministic consequence of the interareal dynamics necessary to produce songlike respiratory patterns (Dima et al., 2018), without any additional mechanisms to produce an explicit temporal prediction. While this type of model has yet to be elaborated into other domains of motor control – though evidence of central-peripheral synchrony has been observed in monkeys performing a joystick task (Mulliken et al., 2008) – it is plausible that the central neural activity driving movement and activity from which peripheral event times could be "decoded" are, in fact, the same dynamical trajectory of population activity observed at different timepoints. Interestingly, such a possibility aligns closely with nearly 200 year old ideomotor views of motor control in psychology, in which motor commands and their immediate consequences have been argued to share a common representational medium (Hommel et al., 2001).

Such a view would dissolve the need for, in the case of experiences of thought insertion in schizophrenia, a clear distinction between the initiation of a thought, the predicted time

of the thought, and the actual thought. The trajectories of population-level neural activity underlying all of these could occur synchronously across involved brain areas, and temporal incoherence stemming from a pathological disruption to ordinary interareal dynamics could result in a thought being subjectively experienced as coming from outside oneself – as its neural antecedents have literally been decoupled from the activity by they were actually caused. In fact, the general idea that temporal coherence between central brain regions and physical peripheral events could be involved in generating the subjective experience of agency already has some empirical support; experimental evidence suggests agency judgments for a brain-machine interface depends on the phase of the alpha rhythm in motor cortex and supplementary motor area at the onset of (prosthetic) peripheral movement (Bertoni et al., 2023). If it is indeed the case that temporal coherence between anticipatorily-synchronized motoric neural activity and sensory feedback impinges upon the conscious sense of agency, and this temporal coherence is a determinstic consequence of interareal dynamics when required to successfully produce movement (Dima et al., 2018; Alonso et al., 2015), the implications for self-awareness would be profound. While the sense of agency has largely been framed as arising from an inferential process that occurs on top of motor control (Haggard, 2008, 2017; Synofzik et al., 2008; Seth et al., 2012), some even going so far as to refer to the sense of an agentic self as hallucinatory (Seth, 2021), this view would suggest that a conscious experience of volition is a direct consequence of fluently self-directing movement.

The analytic tools of computational cognitive neuroscience still have great potential for relating specific categories of representations, such as subtypes of motor representations, to subjective experiences such as agency. For this potential to be fully realized, however, the "decodeable" information content of neural recordings must be clearly disambiguated from the representational content of neural systems, which are not directly reflected in the state of a circuit but in the dynamics which constrain how states can evolve over time. To accomplish this disambiguation, it will be critical to leverage insights from theoretical neuroscience to

make predictions about the dynamics which could subserve a given representation, rather than attempting to "find" that representation reflected in neural activity directly. For example, one research team cleverly derived the macroscopic patterns of activity that should be expected from a population of grid cells during spatial navigation and successfully measured that predicted signal in humans with fMRI (Doeller et al., 2010). I suspect there is hope for more data-driven discovery approaches as well. For instance, a simple recurrent neural network (where dynamics depend on internal states) trained to produce realistic patterns of muscle activity displays quasi-rotational trajectories of its internal states similar to motor cortex (Kaufman et al., 2016); likely, then, incorporating an internal state into the artificial agent model used in our "inverse dynamics" localizer would improve prediction of motor cortex substantially.

The development of algorithmic models for musculoskeletal control which robustly produce human-like motor behavior in biomechanically-detailed physics simulations is a promising step toward explaining human motor control (Chiappa et al., 2024). But given that multiple possible physical implementations could instantiate the same algorithm, or even multiple possible algorithms could be employed in different contexts, representational content is unlikely to map onto implementing neural activity patterns in a linear, one-to-one manner (Veillette et al., 2025a). The goal of relating algorithmic representations of motor control and experiences of motor awareness through their neural substrate, however, may still be achievable within a framework that considers representations as dynamical constraints on the evolution of brain states, rather than conflating them to the states themselves. In such a view, a neural population does not *encode* representations but rather *generates* representations; minding this distinction may ultimately dissolve the apparent separation between the mechanistic process that produces volitional actions and the process of demarcating self-actions as volitional in awareness.

REFERENCES

Aarts, H. and van den Bos, K. (2011). On the Foundations of Beliefs in Free Will: Intentional Binding and Unconscious Priming in Self-Agency. *Psychological Science*, 22(4):532–537. Publisher: SAGE Publications Inc.

Abdulkarim, Z., Guterstam, A., Hayatou, Z., and Ehrsson, H. H. (2023). Neural Substrates of Body Ownership and Agency during Voluntary Movement. *Journal of Neuroscience*, 43(13):2362–2380. Publisher: Society for Neuroscience Section: Research Articles.

Alonso, R. G., Trevisan, M. A., Amador, A., Goller, F., and Mindlin, G. B. (2015). A circular model for song motor control in Serinus canaria. *Frontiers in Computational Neuroscience*, 9. Publisher: Frontiers.

Amador, A. and Mindlin, G. B. (2014). Low dimensional dynamics in birdsong production. *The European Physical Journal B*, 87:1–8. ISBN: 1434-6028 Publisher: Springer.

Amador, A., Perl, Y. S., Mindlin, G. B., and Margoliash, D. (2013). Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature*, 495(7439):59–64. Publisher: Nature Publishing Group.

Appelhoff, S., Hurst, A., Lawrence, A., Li, A., Mantilla, R., Yorguin, J., O'Reilly, C., Xiang, L., and Dancker, J. (2022). PyPREP: A Python implementation of the preprocessing pipeline (PREP) for EEG data.

Asai, T., Kanayama, N., Imaizumi, S., Koyama, S., and Kaganoi, S. (2016). Development of Embodied Sense of Self Scale (ESSS): Exploring Everyday Experiences Induced by Anomalous Self-Representation. *Frontiers in Psychology*, 7.

Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4):948–953. Publisher: Proceedings of the National Academy of Sciences.

Bart, V. K. E., Wenke, D., and Rieger, M. (2023). A German translation and validation of the sense of agency scale. *Frontiers in Psychology*, 14:1199648.

Behroozmand, R., Korzyukov, O., Sattler, L., and Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control. *The Journal of the Acoustical Society of America*, 132(4):2468–2477.

Bertoni, T., Noel, J.-P., Bockbrader, M., Foglia, C., Colachis, S., Orset, B., Rezai, A., Panzeri, S., Becchio, C., and Blanke, O. (2023). Pre-movement sensorimotor oscillations shape the sense of agency by gating cortical connectivity.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., and Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20(28):1–6.

Blakemore, S.-J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *NeuroReport*, 11(11):R11.

Boari, S., Perl, Y. S., Amador, A., Margoliash, D., and Mindlin, G. B. (2015). Automatic reconstruction of physiological gestures used in a model of birdsong production. *Journal of Neurophysiology*, 114(5):2912–2922. Publisher: American Physiological Society.

Boudewyn, M. A., Luck, S. J., Farrens, J. L., and Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, 55(6):e13049. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13049.

Bradshaw, A. R., Lametti, D. R., and McGettigan, C. (2021). The Role of Sensory Feedback in Developmental Stuttering: A Review. *Neurobiology of Language*, 2(2):308–334.

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., Resnik, P., and Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, 12:e85012. Publisher: eLife Sciences Publications, Ltd.

Burak, Y. and Fiete, I. R. (2009). Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Computational Biology*, 5(2):e1000291. Publisher: Public Library of Science.

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., and Martin, O. A. (2022). Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103:1–29.

Chambon, V., Filevich, E., and Haggard, P. (2014). What is the Human Sense of Agency, and is it Metacognitive? In Fleming, S. M. and Frith, C. D., editors, *The Cognitive Neuroscience of Metacognition*, pages 321–342. Springer, Berlin, Heidelberg.

Charalampaki, A., Ninija Karabanov, A., Ritterband-Rosenbaum, A., Bo Nielsen, J., Roman Siebner, H., and Schram Christensen, M. (2022). Sense of agency as synecdoche: Multiple neurobiological mechanisms may underlie the phenomenon summarized as sense of agency. *Consciousness and Cognition*, 101:103307.

Charles, L., King, J.-R., and Dehaene, S. (2014). Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli. *Journal of Neuroscience*, 34(4):1158–1170. Publisher: Society for Neuroscience Section: Articles.

Chiappa, A. S., Tano, P., Patel, N., Ingster, A., Pouget, A., and Mathis, A. (2024). Acquiring musculoskeletal skills with curriculum-based reinforcement learning. *Neuron*, 112(23):3969–3983.e5.

Christensen, M. S. and Grünbaum, T. (2018). Sense of agency for movements. *Consciousness and Cognition*, 65:27–47.

Chung, S. and Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144.

Churchill, N. W., Spring, R., Grady, C., Cimprich, B., Askren, M. K., Reuter-Lorenz, P. A., Jung, M. S., Peltier, S., Strother, S. C., and Berman, M. G. (2016). The suppression of scale-free fMRI brain dynamics across three different sources of effort: aging, task novelty and task difficulty. *Scientific Reports*, 6(1):30895. Number: 1 Publisher: Nature Publishing Group.

Churchland, M. M. and Shenoy, K. V. (2024). Preparatory activity and the expansive null-space. *Nature Reviews Neuroscience*, 25(4):213–236. ISBN: 1471-003X Publisher: Nature Publishing Group UK London.

Constant, M., Salomon, R., and Filevich, E. (2022). Judgments of agency are affected by sensory noise without recruiting metacognitive processing. *eLife*, 11:e72356. Publisher: eLife Sciences Publications, Ltd.

Crapse, T. B. and Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600. Publisher: Nature Publishing Group.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, 10:604.

Daou, A. and Margoliash, D. (2020). Intrinsic neuronal properties represent song and error in zebra finch vocal learning. *Nature Communications*, 11(1):952. Publisher: Nature Publishing Group.

David, N. (2012). New frontiers in the neuroscience of the sense of agency. *Frontiers in Human Neuroscience*, 6. Publisher: Frontiers.

Dehaene, S. and Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, 70(2):200–227.

Dehaene, S. and King, J.-R. (2016). Decoding the Dynamics of Conscious Perception: The Temporal Generalization Method. In Buzsáki, G. and Christen, Y., editors, *Micro-, Meso- and Macro-Dynamics of the Brain*. Springer, Cham (CH).

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., and Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, 132(9):2531–2540.

Dennett, D. (1993). *Consciousness explained*. Penguin uk.

Desantis, A., Roussel, C., and Waszak, F. (2011). On the influence of causal beliefs on the feeling of agency. *Consciousness and Cognition*, 20(4):1211–1220.

Dewey, J. A. and Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PLoS ONE*, 9. Place: US Publisher: Public Library of Science.

Dharmaprani, D., Nguyen, H. K., Lewis, T. W., DeLosAngeles, D., Willoughby, J. O., and Pope, K. J. (2016). A comparison of independent component analysis algorithms and measures to discriminate between EEG and artifact components. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 825–828. IEEE.

Di Plinio, S., Arnò, S., and Ebisch, S. J. H. (2024). The state-trait sense of self inventory: A psychometric study of self-experience and its relation to psychosis-like manifestations. *Consciousness and Cognition*, 118:103634.

Dima, G. C., Copelli, M., and Mindlin, G. B. (2018). Anticipated synchronization and zero-lag phases in population neural models. *International Journal of Bifurcation and Chaos*, 28(08):1830025. ISBN: 0218-1274 Publisher: World Scientific.

Doeller, C. F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, 463(7281):657–661. Publisher: Nature Publishing Group.

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., and Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.

Eke, A., Herman, P., Kocsis, L., and Kozak, L. (2002). Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement*, 23(1):R1–R38.

Eke, A., Hermán, P., Bassingthwaighte, J., Raymond, G., Percival, D., Cannon, M., Balla, I., and Ikrényi, C. (2000). Physiological time series: Distinguishing fractal noises from motions. *Pflugers Archiv European Journal of Physiology*, 439(4):403–415.

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America*, 70(1):45–50.

Feinberg, I. (1978). Efference Copy and Corollary Discharge: Implications for Thinking and Its Disorders. *Schizophrenia Bulletin*, 4(4):636–640. Place: US Publisher: National Institute of Mental Health.

Fernández, E. S. and Fernández, H. V. (2022). Neuro-Rights and Ethical Ecosystem: The Chilean Legislation Attempt. In López-Silva, P. and Valera, L., editors, *Protecting the Mind: Challenges in Law, Neuroprotection, and Neurorights*, pages 129–137. Springer International Publishing, Cham.

Fetterman, G. C. and Margoliash, D. (2023). Rhythmically bursting songbird vocomotor neurons are organized into multiple sequences, suggesting a network/intrinsic properties model encoding song and error, not time. Pages: 2023.01.23.525213 Section: New Results.

Fiedler, K., Schott, M., and Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6):1231–1236.

Fiete, I. R., Burak, Y., and Brookings, T. (2008). What Grid Cells Convey about Rat Location. *Journal of Neuroscience*, 28(27):6858–6871. Publisher: Society for Neuroscience Section: Articles.

Fitzgibbon, S. P., Lewis, T. W., Powers, D. M. W., Whitham, E. W., Willoughby, J. O., and Pope, K. J. (2013). Surface Laplacian of Central Scalp Electrical Signals is Insensitive to Muscle Contamination. *IEEE Transactions on Biomedical Engineering*, 60(1):4–9. Conference Name: IEEE Transactions on Biomedical Engineering.

Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8.

Frith, C. (1987). The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. *Psychological Medicine*, 17(3):631–648. Publisher: Cambridge University Press.

Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and Cognition*, 21(1):52–54.

Galang, C. M., Malik, R., Kinley, I., and Obhi, S. S. (2021). Studying sense of agency online: Can intentional binding be observed in uncontrolled online settings? *Consciousness and Cognition*, 95:103217.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, 4(1):14–21.

Gallagher, S. (2012). Multiple aspects in the sense of agency. *New ideas in psychology*, 30(1):15–31. ISBN: 0732-118X Publisher: Elsevier.

Galvin, S. J., Podd, J. V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, 10(4):843–876.

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., and Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128. Publisher: Nature Publishing Group.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390.

Gneiting, T. and Schlather, M. (2004). Stochastic models which separate fractal dimension and Hurst effect. *SIAM Review*, 46(2):269–282. arXiv:physics/0109031.

Golfinopoulos, E., Tourville, J. A., and Guenther, F. H. (2010). The integration of large-scale neural network modeling and functional brain imaging in speech motor control. *NeuroImage*, 52(3):862–874.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3):280–301.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, 9(12):934–946. Number: 12 Publisher: Nature Publishing Group.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4):196–207. Number: 4 Publisher: Nature Publishing Group.

Haggard, P. and Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, 12(4):695–707.

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4):382–385. Number: 4 Publisher: Nature Publishing Group.

Hall, L., Dawel, A., Greenwood, L.-M., Monaghan, C., Berryman, K., and Jack, B. N. (2023). Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components. *Psychophysiology*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14363.

Hall, L., Johansson, P., Tärning, B., Sikström, S., and Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, 117(1):54–61.

Harutyunyan, A., Dabney, W., Mesnard, T., Gheshlaghi Azar, M., Piot, B., Heess, N., van Hasselt, H. P., Wayne, G., Singh, S., Precup, D., and Munos, R. (2019). Hindsight Credit Assignment. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

Heald, J. B., Lengyel, M., and Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 600(7889):489–493. Publisher: Nature Publishing Group.

Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283.

Hirsh, I. J. and Watson, C. S. (1996). Auditory psychophysics and perception. *Annual review of psychology*, 47(1):461–484. ISBN: 0066-4308 Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24(5):849–878.

Howell, P. and Archer, A. (1984). Susceptibility to the effects of delayed auditory feedback. *Perception & Psychophysics*, 36(3):296–302.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J physiol*, 148(3):574–591.

Hurault, J.-C., Broc, G., Crône, L., Tedesco, A., and Brunel, L. (2020). Measuring the Sense of Agency: A French Adaptation and Validation of the Sense of Agency Scale (F-SoAS). *Frontiers in Psychology*, 11.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458. Publisher: Nature Publishing Group.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634. Conference Name: IEEE Transactions on Neural Networks.

Ibáñez-Molina, A. J. and Iglesias-Parro, S. (2014). Fractal characterization of internally and externally generated conscious experiences. *Brain and Cognition*, 87:69–75.

Imaizumi, S. and Tanno, Y. (2019). Intentional binding coincides with explicit sense of agency. *Consciousness and Cognition*, 67:1–15.

Issa, M. F., Khan, I., Ruzzoli, M., Molinaro, N., and Lizarazu, M. (2024). On the speech envelope in the cortical tracking of speech. *NeuroImage*, 297:120675.

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., and Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429.

Jenkins, A. H. (2001). Individuality in Cultural Context: The Case for Psychological Agency. *Theory & Psychology*, 11(3):347–362. Publisher: SAGE Publications Ltd.

Jensen, K. M. and MacDonald, J. A. (2023). Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components. *Psychophysiology*, 60(7):e14245. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14245.

Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science*, 310(5745):116–119. Publisher: American Association for the Advancement of Science.

Joris, P. X., Smith, P. H., and Yin, T. C. T. (1998). Coincidence Detection in the Auditory System: 50 Years after Jeffress. *Neuron*, 21(6):1235–1238. Publisher: Elsevier.

Kakade, S. M. (2001). A Natural Policy Gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Kang, S. Y., Im, C.-H., Shim, M., Nahab, F. B., Park, J., Kim, D.-W., Kakareka, J., Miletta, N., and Hallett, M. (2015). Brain Networks Responsible for Sense of Agency: An EEG Study. *PLOS ONE*, 10(8):e0135261. Publisher: Public Library of Science.

Kardan, O., Adam, K. C. S., Mance, I., Churchill, N. W., Vogel, E. K., and Berman, M. G. (2020). Distinguishing cognitive effort and working memory load using scale-invariance and alpha suppression in EEG. *NeuroImage*, 211:116622.

Kasahara, S., Nishida, J., and Lopes, P. (2019). Preemptive action: Accelerating human reaction using electrical muscle stimulation without compromising agency. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Kasahara, S., Takada, K., Nishida, J., Shibata, K., Shimojo, S., and Lopes, P. (2021). Preserving Agency During Electrical Muscle Stimulation Training Speeds up Reaction Time Directly After Removing EMS. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Kaufman, M. T., Seely, J. S., Sussillo, D., Ryu, S. I., Shenoy, K. V., and Churchland, M. M. (2016). The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type. *eNeuro*, 3(4). Publisher: Society for Neuroscience Section: New Research.

Kayser, J. and Tenke, C. E. (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology. *International Journal of Psychophysiology*, 97(3):171–173.

Kesić, S. and Spasić, S. Z. (2016). Application of Higuchi's fractal dimension from basic to clinical neurophysiology: A review. *Computer Methods and Programs in Biomedicine*, 133:55–70.

Kim, K. S., Wang, H., and Max, L. (2020). It's About Time: Minimizing Hardware and Software Latencies in Speech Research With Real-Time Auditory Feedback. *Journal of Speech, Language, and Hearing Research*, 63(8):2522–2534. Publisher: American Speech-Language-Hearing Association.

King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

Kragel, P. A., Koban, L., Barrett, L. F., and Wager, T. D. (2018). Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron*, 99(2):257–273. Publisher: Elsevier.

Kriegeskorte, N. and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160. Publisher: Nature Publishing Group.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868. Publisher: Proceedings of the National Academy of Sciences.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412. Publisher: Elsevier.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. Publisher: Frontiers.

Kruse, E., Lesh, T., Board, S., Carter, C., and Joiner, W. (2022). P588. Exploring the Neural Correlates of Sense of Agency Deficits in Psychosis: A DTI Study. *Biological Psychiatry*, 91(9):S327–S328. Publisher: Elsevier.

Kühn, S., Nenchev, I., Haggard, P., Brass, M., Gallinat, J., and Voss, M. (2011). Whodunnit? Electrophysiological Correlates of Agency Judgements. *PLOS ONE*, 6(12):e28657. Publisher: Public Library of Science.

Langdon, C., Genkin, M., and Engel, T. A. (2023). A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*, 24(6):363–377. Publisher: Nature Publishing Group.

Legaspi, R. and Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nature Communications*, 10(1):4250. Number: 1 Publisher: Nature Publishing Group.

Leibfried, F., Pascual-Diaz, S., and Grau-Moya, J. (2019). A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32.

Lhermitte, F., Pillon, B., and Serdaru, M. (1986). Human autonomy and the frontal lobes. Part I: Imitation and utilization behavior: A neuropsychological study of 75 patients. *Annals of Neurology*, 19(4):326–334. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.410190404.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Libet, B., Alberts, W. W., Wright, E. W., and Feinstein, B. (1967). Responses of Human Somatosensory Cortex to Stimuli below Threshold for Conscious Sensation. *Science*, 158(3808):1597–1600. Publisher: American Association for the Advancement of Science.

Liljeholm, M. (2022). Flexible control as surrogate reward or dynamic reward maximization. *Cognition*, 229:105262.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., and Johansson, P. (2014). Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say. *Psychological Science*, 25(6):1198–1205. Publisher: SAGE Publications Inc.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, second edition.* MIT Press.

Lush, P., Caspar, E. A., Cleeremans, A., Haggard, P., Magalhães De Saldanha da Gama, P. A., and Dienes, Z. (2017). The Power of Suggestion: Posthypnotically Induced Changes in the Temporal Binding of Intentional Action Outcomes. *Psychological Science*, 28(5):661–669. Publisher: SAGE Publications Inc.

Lush, P., Roseboom, W., Cleeremans, A., Scott, R. B., Seth, A. K., and Dienes, Z. (2019). Intentional binding as Bayesian cue combination: Testing predictions with trait individual differences. *Journal of Experimental Psychology: Human Perception and Performance*, 45(9):1206. ISBN: 1939-1277 Publisher: American Psychological Association.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information.* MIT press.

Marshall, N. J., Glaser, J. I., Trautmann, E. M., Amematsro, E. A., Perkins, S. M., Shadlen, M. N., Abbott, L. F., Cunningham, J. P., and Churchland, M. M. (2022). Flexible neural control of motor units. *Nature neuroscience*, 25(11):1492–1504. ISBN: 1097-6256 Publisher: Nature Publishing Group US New York.

Martis, R. J., Tan, J. H., Chua, C. K., Loon, T. C., Yeo, S. W. J., and Tong, L. (2015). Epileptic eeg classification using nonlinear parameters on different frequency bands. *Journal of Mechanics in Medicine and Biology*, 15(03):1550040. Publisher: World Scientific Publishing Co.

Matell, M. S. and Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cognitive Brain Research*, 21(2):139–170.

Mathis, M. W., Rotondo, A. P., Chang, E. F., Tolias, A. S., and Mathis, A. (2024). Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832. Publisher: Elsevier.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., and Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046. Publisher: Nature Publishing Group.

Mindlin, G. (2013). The physics of birdsong production. *Contemporary Physics*, 54(2):91–96. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00107514.2013.810852.

Moll, F. W., Kranz, D., Corredera Asensio, A., Elmaleh, M., Ackert-Smith, L. A., and Long, M. A. (2023). Thalamus drives vocal onsets in the zebra finch courtship song. *Nature*, 616(7955):132–136. Publisher: Nature Publishing Group.

Moore, J. and Haggard, P. (2006). Commentary on 'How something can be said about telling more than we can know: On choice blindness and introspection'. *Consciousness and Cognition*, 15(4):693–696. ISBN: 1090-2376.

Mulliken, G. H., Musallam, S., and Andersen, R. A. (2008). Forward estimation of movement state in posterior parietal cortex. *Proceedings of the National Academy of Sciences*, 105(24):8170–8177. Publisher: Proceedings of the National Academy of Sciences.

Nadeau, C. and Bengio, Y. (1999). Inference for the Generalization Error. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.1058.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231. ISBN: 1939-1471 Publisher: American Psychological Association.

Norton, K. G. and Liljeholm, M. (2020). The Rostrolateral Prefrontal Cortex Mediates a Preference for High-Agency Environments. *Journal of Neuroscience*, 40(22):4401–4409. Publisher: Society for Neuroscience Section: Research Articles.

Oganian, Y. and Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, 5(11):eaay6279. Publisher: American Association for the Advancement of Science.

Ohata, R., Asai, T., Kadota, H., Shigemasu, H., Ogawa, K., and Imamizu, H. (2020). Sense of Agency Beyond Sensorimotor Process: Decoding Self-Other Action Attribution in the Human Brain. *Cerebral Cortex*, 30(7):4076–4091.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, 10(5):e1003588. Publisher: Public Library of Science.

Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F., and Fee, M. S. (2015). Growth and splitting of neural sequences in songbird vocal development. *Nature*, 528(7582):352–357. ISBN: 0028-0836 Publisher: Nature Publishing Group UK London.

Oldfield, R. C. (2013). Edinburgh Handedness Inventory. Institution: American Psychological Association.

Oliver, R. and Ballester, J. (1998). Is there memory in solar activity? *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 58(5):5650–5654.

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, 13(1):1–30.

Panikkath, R., Panikkath, D., Mojumder, D., and Nugent, K. (2014). The alien hand syndrome. *Proceedings (Baylor University. Medical Center)*, 27(3):219–220.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of statistical software*, 35(4):1–81.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Perl, Y. S., Arneodo, E. M., Amador, A., Goller, F., and Mindlin, G. B. (2011). Reconstruction of physiological instructions from Zebra finch song. *Physical Review E*, 84(5):051909. Publisher: American Physical Society.

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., and Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, 6(1):103. Publisher: Nature Publishing Group.

Press, C., Thomas, E. R., and Yon, D. (2023). Cancelling cancellation? Sensorimotor control, agency, and prediction. *Neuroscience & Biobehavioral Reviews*, 145:105012.

Qian, B. and Rasheed, K. (2004). Hurst exponent and financial market predictability. In *Proceedings of the IASTED International Conference*, Cambridge, MA.

Ramachandran, V. S. and Hirstein, W. (1998). The perception of phantom limbs. The D. O. Hebb lecture. *Brain*, 121(9):1603–1630.

Render, A. and Jansen, P. (2021). Influence of arousal on intentional binding: Impaired action binding, intact outcome binding. *Attention, Perception, & Psychophysics*, 83(1):103–113.

Rigoni, D., Kühn, S., Sartori, G., and Brass, M. (2011). Inducing Disbelief in Free Will Alters Brain Correlates of Preconscious Motor Preparation: The Brain Minds Whether We Believe in Free Will or Not. *Psychological Science*, 22(5):613–618. Publisher: SAGE Publications Inc.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-Resolutions Inference for brain imaging. *NeuroImage*, 181:786–796.

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Ruiz de Miras, J., Soler, F., Iglesias-Parro, S., Ibáñez-Molina, A. J., Casali, A. G., Laureys, S., Massimini, M., Esteban, F. J., Navas, J., and Langa, J. A. (2019). Fractal dimension analysis of states of consciousness and unconsciousness using transcranial magnetic stimulation. *Computer Methods and Programs in Biomedicine*, 175:129–137.

Russo, A. A., Khajeh, R., Bittner, S. R., Perkins, S. M., Cunningham, J. P., Abbott, L. F., and Churchland, M. M. (2020). Neural Trajectories in the Supplementary Motor Area and Motor Cortex Exhibit Distinct Geometries, Compatible with Different Classes of Computation. *Neuron*, 107(4):745–758.e6.

Sassenhagen, J. and Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, 56(6):e13335. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13335.

Saxena, S., Russo, A. A., Cunningham, J., and Churchland, M. M. (2022). Motor cortex activity across movement speeds is predicted by network-level strategies for generating muscle activity. *Elife*, 11:e67620. ISBN: 2050-084X Publisher: eLife Sciences Publications Limited.

Schneider, D. M. and Mooney, R. (2018). How Movement Modulates Hearing. *Annual Review of Neuroscience*, 41(Volume 41, 2018):553–572. Publisher: Annual Reviews.

Schneider, D. M., Sundararajan, J., and Mooney, R. (2018). A cortical filter that learns to suppress the acoustic consequences of movement. *Nature*, 561(7723):391–395. Publisher: Nature Publishing Group.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. pages 92–96, Austin, Texas.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, 8(10):1391–1400. Number: 10 Publisher: Nature Publishing Group.

Seth, A. (2021). *Being You: A New Science of Consciousness*. Penguin. Google-Books-ID: arVCEAAAQBAJ.

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An Interoceptive Predictive Coding Model of Conscious Presence. *Frontiers in Psychology*, 2. Publisher: Frontiers.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sharpee, T. O., Nagel, K. I., and Doupe, A. J. (2011). Two-dimensional adaptation in the auditory forebrain. *Journal of Neurophysiology*, 106(4):1841–1861. Publisher: American Physiological Society.

Sidarus, N. and Haggard, P. (2016). Difficult action decisions reduce the sense of agency: A study using the Eriksen flanker task. *Acta Psychologica*, 166:1–11.

Simonyan, K. and Horwitz, B. (2011). Laryngeal Motor Cortex and Control of Speech in Humans. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 17(2):197–208.

Sitt, J. D., Amador, A., Goller, F., and Mindlin, G. B. (2008). Dynamical origin of spectrally rich vocalizations in birdsong. *Physical Review E*, 78(1):011905. Publisher: American Physical Society.

Smith, N. J. and Kutas, M. (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2):169–181. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.12320.

Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98.

Sokunbi, M. O., Gradin, V. B., Waiter, G. D., Cameron, G. G., Ahearn, T. S., Murray, A. D., Steele, D. J., and Staff, R. T. (2014). Nonlinear Complexity Analysis of Brain fMRI Signals in Schizophrenia. *PLOS ONE*, 9(5):e95146. Publisher: Public Library of Science.

Sreenivasan, S. and Fiete, I. (2011). Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337. Publisher: Nature Publishing Group.

Stevens, D. A. and O'Connell, R. J. (1991). Individual differences in thresholds and quality reports of human subjects to various odors. *Chemical Senses*, 16(1):57–67.

Stier, A., Cardenas-Iniguez, C., Kardan, O., Moore, T., Meyer, F., Rosenberg, M., Kaczkurkin, A., Lahey, B., and Berman, M. (2021). A Scale-Free Gradient of Cognitive Resource Disruptions in Childhood Psychopathology. Technical report.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Suzuki, K., Lush, P., Seth, A. K., and Roseboom, W. (2019). Intentional Binding Without Intentional Action. *Psychological Science*, 30(6):842–853. Publisher: SAGE Publications Inc.

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, 17(1):219–239.

Tajima, D., Nishida, J., Lopes, P., and Kasahara, S. (2022). Whose Touch is This?: Understanding the Agency Trade-Off Between User-Driven Touch vs. Computer-Driven Touch. *ACM Transactions on Computer-Human Interaction*, 29(3):24:1–24:27.

Tanaka, T., Matsumoto, T., Hayashi, S., Takagi, S., and Kawabata, H. (2019). What Makes Action and Outcome Temporally Close to Each Other: A Systematic Review and Meta-Analysis of Temporal Binding. *Timing & Time Perception*, 7(3):189–218. Publisher: Brill.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866. Publisher: Nature Publishing Group.

Tapal, A., Oren, E., Dar, R., and Eitam, B. (2017). The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. *Frontiers in Psychology*, 8.

Tellegen, A. and Atkinson, G. (1981). Tellegen Absorption Scale. *Journal of Abnormal Psychology*.

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., and Gallant, J. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3):289–316. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/net.12.3.289.316.

Timm, J., Schönwiesner, M., Schröger, E., and SanMiguel, I. (2016). Sensory suppression of brain responses to self-generated sounds is observed with and without the perception of agency. *Cortex*, 80:5–20.

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America*, 83(4):1536–1552. ISBN: 0001-4966 Publisher: Acoustical Society of America.

Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981. Publisher: Routledge _eprint: https://doi.org/10.1080/01690960903498424.

Tsakiris, M., Longo, M. R., and Haggard, P. (2010). Having a body versus moving your body: Neural signatures of agency and body-ownership. *Neuropsychologia*, 48(9):2740–2749.

Tsakiris, M., Prabhu, G., and Haggard, P. (2006). Having a body versus moving your body: How agency structures body-ownership. *Consciousness and Cognition*, 15(2):423–432.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432. arXiv:1507.04544 [stat].

Veillette, J. P., Chao, A. F., Nith, R., Lopes, P., and Nusbaum, H. C. (2025a). Overlapping Cortical Substrate of Biomechanical Control and Subjective Agency. *Journal of Neuroscience*. ISBN: 0270-6474 Publisher: Society for Neuroscience.

Veillette, J. P., Ho, L., and Nusbaum, H. C. (2023a). Permutation-based group sequential analyses for cognitive neuroscience. *NeuroImage*, 277:120232.

Veillette, J. P., Ho, L., and Nusbaum, H. C. (2024). Metacognition bridges experiences and beliefs in sense of agency. *Consciousness and Cognition*, 124:103745.

Veillette, J. P., Lopes, P., and Nusbaum, H. C. (2023b). Temporal Dynamics of Brain Activity Predicting Sense of Agency over Muscle Movements. *Journal of Neuroscience*, 43(46):7842–7852. Publisher: Society for Neuroscience Section: Research Articles.

Veillette, J. P. and Nusbaum, H. C. (2025). Bayesian p-curve mixture models as a tool to dissociate effect size and effect prevalence. *Communications Psychology*, 3(1):9. ISBN: 2731-9121 Publisher: Nature Publishing Group UK London.

Veillette, J. P., Rosen, J., Margoliash, D., and Nusbaum, H. C. (2025b). Timing of speech in brain and glottis and the feedback delay problem in motor control.

Voss, M., Ingram, J. N., Haggard, P., and Wolpert, D. M. (2006). Sensorimotor attenuation by central motor command signals in the absence of movement. *Nature Neuroscience*, 9(1):26–27. Number: 1 Publisher: Nature Publishing Group.

Wang, S., Rajananda, S., Lau, H., and Knotts, J. D. (2020). New measures of agency from an adaptive sensorimotor task. *PLOS ONE*, 15(12):e0244113. Publisher: Public Library of Science.

Wegner, D. M. (2017). *The illusion of conscious will*. MIT press.

Wen, W., Charles, L., and Haggard, P. (2023). Metacognition and sense of agency. *Cognition*, 241:105622.

Wen, W. and Haggard, P. (2018). Control Changes the Way We Look at the World. *Journal of Cognitive Neuroscience*, 30(4):603–619.

Wen, W. and Haggard, P. (2020). Prediction error and regularity detection underlie two dissociable mechanisms for computing the sense of agency. *Cognition*, 195:104074.

Westfall, J. (2017). Statistical details of the default priors in the Bambi library. Technical report. Publication Title: arXiv e-prints ADS Bibcode: 2017arXiv170201201W.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, 12(1):1–24.

Wolf, S. L., Lecraw, D. E., Barton, L. A., and Jann, B. B. (1989). Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Experimental Neurology*, 104(2):125–132.

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, 12(12):739–751. Number: 12 Publisher: Nature Publishing Group.

Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3(11):1212–1217. Number: 11 Publisher: Nature Publishing Group.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An Internal Model for Sensorimotor Integration. *Science*, 269(5232):1880–1882. Publisher: American Association for the Advancement of Science.

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7):1317–1329.

Yamamoto, K. and Kawabata, H. (2014). Adaptation to delayed auditory feedback induces the temporal recalibration effect in both speech perception and production. *Experimental Brain Research*, 232(12):3707–3718.

Zhuang, C., Meidenbauer, K. L., Kardan, O., Stier, A. J., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., and Berman, M. G. (2022). Scale invariance in fNIRS as a measurement of cognitive load. *Cortex*, 154:62–76.