How brains build higher order representations of uncertainty

Megan A. K. Peters*[1,2,3,4,5], Hojjat Azimi Asrari*[1]

[1]Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92617 USA
[2]Department of Logic & Philosophy of Science, University of California Irvine, Irvine, CA 92617 USA
[3]Center for the Neurobiology of Learning & Memory, University of California Irvine, Irvine, CA 92617 USA
[4]Center for Theoretical Behavioral Sciences, University of California Irvine, Irvine, CA 92617 USA
[5]Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research, Toronto, Ontario, M5G 1M1 Canada

* Corresponding authors: M.A.K.P. (megan.peters@uci.edu) and H.A.A. (hazimias@uci.edu)

Abstract

Higher-order representations (HORs) are neural or computational states that are "about" first-order representations (FORs), encoding information not about the external world per se but about the agent's own representational processes—such as the reliability, source, or structure of a FOR. These HORs appear critical to metacognition, learning, and even consciousness by some accounts, yet their dimensionality, construction, and neural substrates remain poorly understood. Here, we propose that metacognitive estimates of uncertainty or noise reflect a read-out of "posterior-like" HORs from a Bayesian perspective. We then discuss how these posterior-like HORs reflect a combination of "likelihood-like" estimates of current FOR uncertainty and "prior-like" learned distributions over expected FOR uncertainty, and how various emerging engineering and theory-based analytical approaches may be employed to examine the estimation processes and neural correlates associated with these highly under-explored components of our experienced uncertainty.

Keywords: neural representations, higher-order representations, uncertainty, noise, probabilistic population codes, neuroimaging, generative artificial intelligence, reinforcement learning, decoded neurofeedback

Introduction

As comprehensively discussed by Baker and colleagues (Baker et al., 2022), the definition of a 'neural representation' is hotly debated: ask three researchers (or three *fields* of research), and you'll get three different answers (Favela & Machery, 2023, 2025; Machery, 2025; Vilarroya, 2017). Here, we take at face value that one possible definition of neural representations is that they are more than just statistical covariation between patterns of neural response and relevant aspects of an observer's environment or mental processing (Baker et al., 2022; Ritchie et al., 2019), instead reflecting the sorts of *mental structures* that observers use to perceive, reason about, and engage with their environments (Tarr & Vuong, 2002). This leads us to want to explore the *kinds* of representations that might be relevant for an agent's behavior or cognitive capacities, and to then examine how neural correlates of such representations might be scientifically studied.

In this paper, we first examine the difference between the more-commonly studied "first-order representations" (FORs), which encode or represent features of the external world, and the less-commonly studied "higher-order representations" (HORs), which encode properties of FORs such as their noise or evidentiary strength. While HORs are foundational to learning, introspection, and potentially consciousness, their computational structure and neural instantiation remain poorly understood. We then introduce and discuss the proposal that HORs about noise or uncertainty might represent these aspects of their target FORs specifically in multiple variants, following a general Bayesian-like framework: (1) by encoding the current unreliability or variability of a first-order signal, which we term *likelihood-like HORs*; (2) by maintaining structured expectations the typical noise associated with certain stimuli or states, which we term *prior-like HORs*; and (3) by combining these estimates into *posterior-like* HORs about uncertainty. We synthesize recent theoretical and analytical solutions to accessing and characterizing each of these HOR variants, and discuss how they may be integrated to ultimately produce these posterior-like higher-order estimates of FOR uncertainty which may drive metacognitive assessments, learning, or conscious awareness.

First-order versus higher-order (neural) representations

In much of the literature on neural representations, the target of study is representations that are "about" aspects of the agent's environment: objects or features of the environment itself (Baker et al., 2022; Tarr & Vuong, 2002), decision variables about that environment leading to behavioral outputs (Gold & Shadlen, 2007), memories (Squire & Zola-Morgan, 1991), goals (Miller & Cohen, 2001), or actions the agent might take to achieve such goals (Rizzolatti & Craighero, 2004; Thornton & Tamir, 2024), for example. Here, we use "about" in quotes to emphasize that the target of a mental representation – that is, what it refers to (see also the Representational Theory of Mind; Schneider, 2020; Von Eckardt, 2012) – may play a key role in how we design both scientific and philosophical lines of inquiry to characterize that representation, much as we attend to how a model's target constrains the construction of and interpretation of that model in the philosophy of modeling (Elliott-Graves, 2020; Schneider, 2020; Weisberg, 2013). Because these representations are "about" the observer's external environment (or history of its perceptions about and actions on that environment, as in memory), enabling the observer even to run predictive models based on such representations in order to plan and execute goal-directed behaviors (Friston, 2010), the literature often refers to these representations as *first-order representations* (FORs).

But brains do not merely represent features of the external world; they also monitor and reflect the agent's own ongoing processing, mental state, or mental structures – including the organism's own models or representations of the world. These *higher order representations* (HORs) are thus defined as being "about" FORs (Brown et al., 2019; Cleeremans et al., 2007). HORs could for example represent the signal strength in a FOR (regardless of its content) (Fleming, 2020), or whether a FOR's content was likely externally or internally generated (i.e., real or a hallucination (Lau, 2019; Michel, 2024)), or the magnitude of noise or uncertainty present in a FOR (Winter & Peters, 2022). Note that these HORs should be distinguished on this basis from neural representations of aspects of "higher order cognition" such as executive function or task switching, instead referring to representations that are about one's own mental state or ongoing processing.

Such HORs receive somewhat less attention than FORs in the general literature on neural representation, their study being largely confined to those who study metacognition, meta-learning, and similar "thinking about thinking" type approaches (Dunlosky & Metcalfe, 2009; Proust, 2007). One possible reason for this relatively smaller literature is that studying such HORs is methodologically and conceptually challenging

because the processes giving rise to them may not be easily anchorable to objectively measurable observables such as behavioral reports (Peters, 2025). This challenge has been long noted in the literature, perhaps most famously with Nisbett & Wilson's (Nisbett & Wilson, 1977) observation that

> Subjects are sometimes (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response. It is proposed that when people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. This suggests that though people may not be able to observe directly their cognitive processes, they will sometimes be able to report accurately about them. (p. 231)

This unreliability of introspective processes (which give rise to or are supported by HORs) has led some – especially in the consciousness science community – to hypothesize HORs to be so problematic that scientific inquiry into the topic in general may be impossible (Dennett, 1991; Peels, 2016; Schwitzgebel, 2008, 2011). However, others – as in the metacognition community – have taken the stance that HORs and their associated behavioral reports may be unreliable, but that systematic patterns can nevertheless be discovered and characterized through research programs specifically designed to target their underlying processes (e.g., Fleming, 2023; Fleming and Lau, 2014; Kammerer and Frankish, 2023; Peters, 2020, 2022, 2025; Rahnev, 2021). In this piece, we build upon this second, hopeful perspective to explore how a particular kind of HOR might be constructed and scientifically studied: HORs which are about the reliability of an observer's ongoing representations and decision processes.

Why study higher-order representations of uncertainty?

We focus on HORs which are specifically about noise or uncertainty in a FOR for several reasons. First, HORs about noise or uncertainty appear especially relevant for learning. For example, observers who are more "introspectively calibrated" (Fleming & Lau, 2014; Maniscalco, Charles, & Peters, 2024) — i.e., those whose confidence better corresponds with choice accuracy and learned information — tend to learn about their environments more quickly (Frömer et al., 2021; Hainguerlot et al., 2018; Meyniel, Sigman, & Mainen, 2015). This means that observers must calibrate their introspective judgments to reflect on learned environmental variables (Koriat, 1997; Meyniel & Dehaene, 2017; Meyniel, Schlunegger, & Dehaene, 2015) – even in the absence of external feedback – to further guide the learning process itself (Guggenmos, 2022; Guggenmos et al., 2016). Studies and interventions seeking to study or even optimize learning thus may strongly benefit from a better understanding of uncertainty-related HORs.

Second, uncertainty-related HORs are also frequently invoked in inquiry into the brain's ability to distinguish reality from imagination (Fleming & Daw, 2017; Gershman, 2019; Lau, 2019) or generate conscious awareness (Brown, 2015; Cleeremans, 2011; Cleeremans et al., 2019; Fleming, 2020; Lau & Rosenthal, 2011; Michel & Lau, 2021; Rosenthal, 2005). Specifically, Higher Order Theories (HOTs) of consciousness posit that the formation and maintenance of a HOR is both necessary and sufficient for a percept, idea, or feeling being conscious (Rosenthal, 2012). There are several well-described HOT variants (Brown et al., 2019). For example, Higher Order State Space (HOSS) theory (Fleming, 2020) posits that a higher order monitoring mechanism assesses the strength (and potentially reliability) of a FOR, such that if this assessment surpasses a threshold, the contents of the FOR rise into awareness. In Perceptual Reality Monitoring (PRM) theory (Lau, 2019; Michel, 2024), it is assumed that a metacognitive mechanism estimates not only FOR strength but also whether the FOR is likely externally- or internally-sourced – i.e., whether it is likely reflect external signals from the environment, or internally-generated imagery or noise, much like the task of a generative adversarial network (GAN) (Gershman, 2019). If the PRM mechanism fails, tagging a FOR as 'real' when it was internally-generated noise, the result is theorized to be a hallucination with conscious, phenomenal quality. Higher-Order Representation of a Representation (HOROR) theory (Brown, 2015) suggests that the content of a FOR is also present at the HOR level, albeit perhaps "redescribed in a different format" (Brown et al., 2019). A major difference among these HOT variants lies in the dimensionality and dimensions of the HOR: in HOSS, there is a signal HOR dimension (signal strength), while in PRM there are two (signal strength; reality vs. imagination) and in HOROR there are more (signal strength; reality vs. imagination;

FOR content). Arbitrating these theories can benefit from more complete descriptions of HORs, including their dimensions and dimensionality. In other words, we must discover whether and how HORs may encode not only the uncertainty in an FOR, but also its strength, spatiotemporal stability, content, or any other descriptives (Peters, 2022) and how read-outs of (or decision policies applied to) such HORs may drive not only metacognitive judgments but also whether the contents of an FOR are phenomenally conscious or available to behavioral report.

The components of higher-order representations of uncertainty

Much work has sought to evaluate how HORs and the metacognitive (confidence) judgments they produce are constructed, as well as the associated neural correlates (e.g., Fleming, 2024; Fleming and Dolan, 2012; Maniscalco and Lau, 2016b; Peters, 2020, 2022; Rahnev, 2021; Shekhar and Rahnev, 2024). These studies have developed a veritable zoo of potential (neural) computations giving rise to uncertainty-related HORs and confidence judgments, including charting paths forward through targeted empirical studies designed to arbitrate such theories (Rahnev et al., 2022).

However, less effort has been devoted to characterizing the *entire* processing chain constructing HORs and confidence judgments. To be more specific, if one posits that confidence judgments result from a read-out of a HOR, then to explain those confidence judgments completely, one must describe (a) the inputs to the function *generating* the HOR in the first place, (b) the function operating on those inputs, (c) the dimensions and dimensionality of the resulting HOR, and (d) the decision policy applied to the HOR to produce a confidence report. As described by Peters (Peters, 2022), nearly all perceptual metacognition and consciousness literature has confined itself to characterizing (b) from this list, with only a few studies examining deviations from the assumed standard inputs of 'stimulus evidence' (e.g. Mamassian and de Gardelle, 2022, 2024; Winter and Peters, 2022) broadly defined. A full characterization of HORs requires attention to all possible components of the metacognitive evaluation process constructing those HORs (Peters, 2022).

How can we discover the neural patterns associated with each of these components or mechanisms? A possible path forward is to specifically seek HORs of *contributors* to the metacognitive estimation process – the *inputs* to the metacognitive computation as well as its outputs. That is, the metacognitive estimation process may be Bayesian-like, in which a *current* estimate of uncertainty or noise is combined with the system's *prior expectations* for noise under present conditions or contexts (Box 1).[1] This proposal cleanly unifies metacognitive estimation and the construction of HORs of uncertainty with the widely successful 'perception as Bayesian inference' framework (Knill & Richards, 1996): the brain uses Bayesian inference to first construct a posterior estimate over the most likely state of the world given available information and prior expectations (FOR), and then again to build a posterior estimate over the most likely level of uncertainty present in that FOR (HOR) (Figure 1).

[1]In this piece we will retain the Bayesian terminology for the sake of brevity and clarity of argument. But even if one doesn't subscribe to the hypothesis that metacognition involves a Bayes-like process combining current estimates of noise with prior expectations over noise, it is reasonable to argue that a critical factor for an organism trying to evaluate its own uncertainty would before it to have some sort of 'anchor' or benchmark against which to compare a current uncertainty estimate. Essentially, the system needs to be able to ask, "Is the FOR-uncertainty I'm estimating right now large or small *relative to the uncertainty I tend to experience*?" Such a comparison process necessitates the presence of some representation of FOR-uncertainty *distributions*.

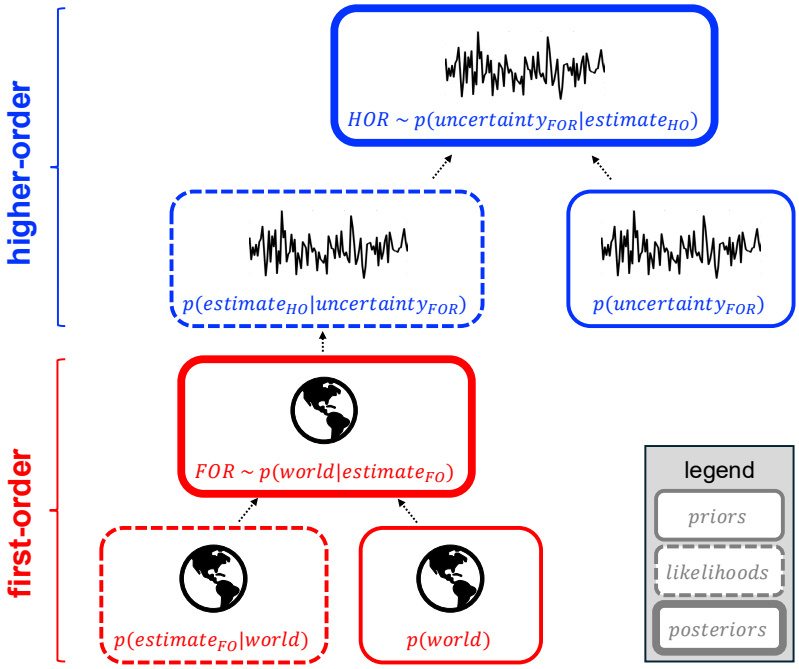

Figure 1: **Visual representation of the proposed hierarchical Bayesian process giving rise to higher-order representations (HORs) of uncertainty.** The system constructs a first-order representation (FOR; red) of the most likely state of the world through Bayesian inference, and then again uses Bayesian inference to construct a higher-order representation (HOR; blue) about the uncertainty present in the FOR.

Box 1: HOR Variants and Bayesian Structure

Higher-order representations (HORs) of uncertainty can be decomposed analogously to components of Bayesian inference:

HOR type	Description
Likelihood-like	Estimates the *momentary reliability* of a current first-order representation (FOR).
Prior-like	Encodes *structured expectations* about the typical noise or reliability of a target FOR along task-relevant dimension(s).
Posterior-like	Integrates both likelihood-like and prior-like components to form a *composite estimate* of FOR uncertainty.

By this formulation, HORs of uncertainty follow the standard Bayesian formulation: $p(uncertainty_{FOR}|estimate_{HO}) \propto p(estimate_{HO}|uncertainty_{FOR})p(uncertainty_{FOR})$.

Recently, Winter & Peters (2022) tested the proposal that the visual system has developed prior expectations over expected uncertainty in FORs as a function of eccentricity across the visual field – parafoveal (central) versus peripheral. They found that simple errors in the distribution of expected uncertainty could explain intriguing dissociations between *actual* uncertainty in FORs (as measured by task performance accuracy) and *estimated* uncertainty (as measured by subjective or metacognitive reports), and how such dissociations could be altered through task manipulations of endogenous attention. Unfortunately, though, most research, to the extent it examines (HO) representations of FOR uncertainty at all, has focused on HORs which reflect either a direct read-out of (FO) uncertainty from the perspective of the experimenter, or the *result* of the estimation process as described above. In the following sections, we explore how both

the *instantaneous estimate* of FOR uncertainty and the *distribution of expected* FOR-noise can also be considered metacognitively constructed HORs of uncertainty – ones which have received almost no attention in the literature.

The "posterior": The experienced FOR uncertainty

As introduced above, essentially all the work in characterizing neural representations of (HO) uncertainty has focused on discovering patterns of neural response which covary with behavioral reports about uncertainty in a task (Walker et al., 2023). Thus, behavioral reports provide an attractive target for revealing posterior-like HORs of uncertainty, since they reflect the result of the brain's own self-monitoring or uncertainty estimation processes. The neural correlates of these posteriors over uncertainty have been comprehensively explored by many others, so we do not extensively discuss these here; interested readers may wish to refer to (Fleming & Dolan, 2012) and (Fleming, 2024) for reviews and discussion.

One can posit many possible estimation processes which may lead to such behavioral outputs (Shekhar & Rahnev, 2024), such as the addition of additional noise or biases at the introspective, self-monitoring level (Boundy-Singer et al., 2023; Mamassian, 2018; Mamassian & de Gardelle, 2022, 2024; Maniscalco, Castaneda, et al., 2024; Maniscalco, Charles, & Peters, 2024; Maniscalco & Lau, 2012, 2014). Most if not all of these implicitly assume that the metacognitive read-out reflects a direct estimate of FOR uncertainty, as with the type 2 noise posited by the signal detection theoretic measure meta-d' (Maniscalco & Lau, 2012, 2014, 2016a) and the 'confidence as a noisy decision reliability estimate' (CASANDRE) model (Boundy-Singer et al., 2023). Yet again, these models do not separate the 'posterior-like' result of this estimation process from the 'likelihood-like' estimate itself, as we begin to explore in this piece.

In addition to studies seeking neural correlates of the posterior-like uncertainty HORs, some model-driven neuroimaging studies have also sought to reveal neural correlates which may arbitrate between the metacognitive computations giving rise to them (e.g., Peters et al., 2017). However, while seeking neural correlates of the results of such estimation processes provides a powerful path towards understanding HORs of FOR-uncertainty, this approach does not offer a concrete focus on the *components* – or *inputs* – to such estimation processes per se (Peters, 2022). These studies are therefore limited in revealing the full heterogeneity or variety of kinds of FOR-uncertainty HORs, thus also limiting visibility into metacognitive computations themselves – specifically, the likelihood-like and prior-like HORs of uncertainty.

The "likelihood": A noisy estimate of current FOR uncertainty

Because behaviorally-reported uncertainty is unlikely to be a direct, noiselessly-perfect readout of FOR uncertainty, instead reflecting the *result* of the (potentially Bayesian) metacognitive uncertainty estimation process (Mamassian, 2024; Winter & Peters, 2022), we would next want to characterize the likelihood function as well as this posterior. We should therefore seek to directly quantify current uncertainty in a target FOR, and then seek neural correlates which may encode this estimate: the likelihood-like HORs of uncertainty.

Engineering approaches designed to measure uncertainty or noise in neural representations may be co-opted in service of this goal, if 'pointed at' FORs. For example, the GLMsingle method (Prince et al., 2022) couples custom hemodynamic response functions (HRFs) with regularization and cross-validation approaches to improve the reliability of beta estimates for single voxels on single trials within a task, which quantify how much a given voxel's activity is predicted by a task-relevent variable. While the goal of GLMsingle is to improve the signal-to-noise ratio of measured BOLD responses via fMRI by discarding the noise, as with any general linear model (GLM) based approach, nuisance regressors are included in the model to explicitly estimate variance *not* associated with the task-relevant variables of interest – i.e., the noise. More recently, a similar approach was developed which leverages generative models to explicitly measure noise distributions in voxel space, termed Generative Modeling of Signal and Noise (GSN) (Kay et al., 2024). Like GLMsingle, the goal is to improve estimates of the signal distribution in BOLD data, in this case by directly estimating the noise and then subtracting it off.

For our purposes, one might be tempted to use GLMsingle or GSN to directly estimate voxel noise and then seek its relationship to 'likelihood-like' HORs of FOR uncertainty, i.e. to discover neural signals or representations which covary with these estimated noise levels. One challenge, though, is that both

GLMsingle and GSN are designed to measure voxel noise rather than FOR noise specifically, meaning that the manner in which they estimate this noise is not at all akin to how the brain might monitor its own FOR noise. In both GLMsingle and GSN, general linear models are coupled with regularization approaches which are unlikely to be directly analogous to any method employed by the brain. These same limitations are unfortunately also true for other methods specifically targeting identifying noise distributions in neural data collected via other methods, such as electrophysiology or calcium imaging (Pospisil & Pillow, 2024; Stringer et al., 2019; Williams & Linderman, 2021). In short, the relationship between voxelwise noise and FOR noise is as complex a the relationship between voxelwise patterns and the mental structures they represent, such that seeking neural correlates of voxelwise noise discovered through GLMsingle or GSN would not necessarily reflect HO estimates of FOR uncertainty per se.

Instead, then, one could "read out" the uncertainty encoded in a (FO) neural representation measured via multi-unit electrophysiology using a model-based approach which explicitly dictates the relationship between neural population responses and the FORs they encode (Walker et al., 2023). One candidate would be probabilistic population codes, which posit that Bayesian uncertainty in a representation is specifically encoded in the gain of neural population responses (Ma & Pouget, 2009; Ma et al., 2006). Noninvasive neuroimaging approaches have also been developed for quantifying uncertainty in FORs based on the probabilistic population coding framework: The TAFKAP method (van Bergen and Jehee, 2021: The Algorithm Formerly Known as PRINCE) and its predecessor PRINCE (van Bergen et al., 2015: **Pr**obabilistic **In**ference from activity in **C**ort**e**x) both directly estimate the (FO) uncertainty in a given neural pattern along a task-relevant dimension by inverting a generative model of stimulus-evoked cortical responses. These methods have been developed to estimate probability distributions reflecting sensory uncertainty in human visual cortex during simple perceptual decision-making tasks, such as estimating the orientation/tilt of an oblique Gabor patch. The authors have reported that they can estimate this FOR uncertainty, and that observers may use knowledge of this uncertainty in their perceptual decisions: higher decoded uncertainty is related to more variable behavioral choices about the stimulus identity, i.e. lower performance and the magnitude of behavioral bias (van Bergen et al., 2015), suggesting that human observers use knowledge of this internal uncertainty in their perceptual decision-making and can monitor fluctuations in this uncertainty from one moment (or trial) to the next.

Importantly, though, the property measured by PRINCE and TAFKAP is FOR uncertainty from one moment to the next rather than the (likelihood-like) HOR about that uncertainty. While the authors claim that observers monitor their own FOR uncertainty and use it in behavioral decisions, their behavioral results demonstrate only that the FOR uncertainty affects decisions – as one would expect from variations in a likelihood causing variations in a posterior judgment when a prior expectation is held constant. Because the authors did not seek to separate the observer's behavioral estimates of FOR uncertainty from actual FOR uncertainty, nor did they measure prior expectations about uncertainty, they could not assess the relationship between measured FOR uncertainty and any (likelihood- or posterior-like) HORs about it. Nevertheless, this method may show promise for future exploration of likelihood-like HOR encoding of FOR uncertainty estimates.

Despite this promising start, however, we also want to note that extending TAFKAP to areas of the brain beyond early visual cortex is also likely to be highly methodologically challenging. The response properties of early visual cortex are extremely well understood: neurons possess orientation selectivity preferences (Brouwer & Heeger, 2011; Haynes & Rees, 2005; Jehee et al., 2012; Kamitani & Tong, 2005; Kay et al., 2008; Serences et al., 2009), and individual neurons' activity exhibit well characterized noise correlations across trials (Goris et al., 2014; Smith & Kohn, 2008). This deep knowledge of visual cortex response properties makes it possible to develop the generative models on which TAFKAP's success relies. Unfortunately, though, response properties of other FORs are less well characterized – for example, selectivity is more mixed in later visual processing areas such as inferiortemporal cortex (e.g., Bao et al., 2020; Chang et al., 2021). Discovering the coding properties of "higher" level FORs beyond early visual cortex is a massive undertaking in its own right. As such response properties are revealed, however, it may be possible to marry TAFKAP-like methods with behavioral metrics of HOR-derived uncertainty estimates.

The "prior": Expectations about FOR uncertainty

As with any Bayesian model, there is one component remaining to discuss: the prior, or the expectations about FOR uncertainty that the observer has built through experience (Series & Seitz, 2013). How might we go about understanding such *expected FOR-noise distribution* neural HORs?

To characterize any novel distribution, one might start by simply taking samples. But we cannot take samples from the posterior (confidence reports or associated neural correlates). To begin measuring *expected noise-distribution* HORs, one might instead employ psychophysical measures such as those previously used to recover priors about environmental variables used in FORs – for example in perception (Adams et al., 2004; Girshick et al., 2011; Odegaard & Shams, 2016; Odegaard et al., 2015; Peters et al., 2015; Series & Seitz, 2013; Stocker & Simoncelli, 2006). In these studies, behavioral measurements are first used to measure the percept (the Bayesian posterior) of e.g. orientation, speed, or object heaviness; then, through manipulating the environmental noise present in the stimuli, one can decompose the combined estimate (posterior distribution) into a Bayesian combination of the instantaneous, noisy estimate of the environmental variable of interest (likelihood) and the prior used by the observer. Such studies have revealed priors across environmental variables such as spatial location (Odegaard et al., 2015), motion speed (Stocker & Simoncelli, 2006), visual contour orientation (Girshick et al., 2011), light source location (Adams et al., 2004), and even tendency to bind multisensory stimuli (Odegaard & Shams, 2016), for example. Used in conjunction with metacognitive judgments about FOR-uncertainty or confidence, this approach may provide insight into task-specific or context-conditioned distributions of expected noise, which could then be used to drive discovery of their neural correlates. However, contextually-conditioned distributions of expected noise then would be confined to a particular variable or task of interest, which – while interesting and fruitful in the context of certain observable variables in the environment such as contour orientation, object density, and so on – will not give us an understanding of the *full* landscape of FOR-noise distributions in the brain, or how they are learned by the system.

Instead of using behavior alone, then, another possibility is to directly sample from the neural prior-like uncertainty HOR itself as it varies across tasks, context, or time using neuroimaging approaches. Note that this approach requires specifically that we sample from the prior-like HOR, not just sample the distribution of noise in the brain across task, context, or time; a resting state scan, for example, would be insufficient. Instead, sampling from the prior-like HOR directly would require identifying a target, task-relevant dimension of the FOR about which uncertainty may be estimated, and being able to track how the brain builds HORs about this dimension so as to eventually map it back to brain response.

We recently proposed an approach to achieving this prior-like HOR sampling: the Noise Estimation through Reinforcement-based Diffusion (NERD) model, which uses generative artificial intelligence (genAI) algorithms designed specifically to learn noise distributions (in service of iteratively 'denoising' patterns – e.g., images, music, or text – to produce a target pattern) from empirical neuroimaging data collected from humans who learned to do a similar task (to iteratively 'denoise' their neural activity patterns to achieve a target pattern) (Azimi Asrari & Peters, 2025; Azimi Azrari & Peters, 2024). After training, NERD possesses in its model architecture and fitted parameters a representation of the *distribution* of voxel pattern noise learned across the task (i.e., $p(noise|step, input_{step})$; Fig. 2a), such that it in essence represents the results of 'sampling the noise'.

To map the learned noise-distribution HORs back to neural activity patterns, we sought a set of sampled points in neural state space which could be directly compared to sampled points in the NERD model. This required that the task learned by NERD, the way NERD learns the task, and the data used to train NERD to do this task, need be as analogous as possible to the task and data structure for the human data. In other words, the model must have learned about its own prior-like noise distributions in a similar way as a target human did, and using similar data.

To accomplish this goal, we used data from a Decoded Neurofeedback (DecNef) (Fig. 2b,c) task, in which human subjects learned to alter the patterns of their own brain response in order to achieve a target goal pattern. Functional magnetic resonance imaging (fMRI) DecNef combines real-time fMRI with multivariate pattern analysis to allow individuals to regulate complex brain activity patterns voluntarily (Cortese et al., 2021; LaConte, 2011). Machine learning algorithms are trained to decode specific (FO) mental states from participants' fMRI activation patterns; by providing continuous feedback on their ongoing mental state, DecNef trains participants to modulate the associated brain activity patterns in target regions of interest

(ROIs) (Shibata et al., 2011; Watanabe et al., 2017), specifically along a task-relevant dimension.

We proposed that one way human subjects can solve the DecNef task is by learning about the uncertainty in their own neural representations, and then navigating this distribution of noise – essentially 'denoising' neural patterns – to achieve their target goal (Azimi Asrari & Peters, 2025; Azimi Azrari & Peters, 2024). We posited this in part because in DecNef the goal state is entirely unknown to the subject, and the subject receives no explicit instructions on what the target pattern should be. It has previously been proposed that human subjects learn to achieve target patterns through reinforcement learning (RL), because the DecNef procedure involves a pretrained classifier comparing the current brain state to the target brain state and then displaying the discrepancy to the user in the form of visual feedback reward (Shibata et al., 2011). Because no instructions are given to the subject other than 'maximize the reward you can achieve', we hypothesized that the brain may 'solve' DecNef through engaging a procedure it *does* know how to do: uncertainty reduction. Uncertainty reduction is a core capacity for all biological brains which may guide perception, action, curiosity, and information seeking (De Ridder et al., 2014; Friston et al., 2017; Gottlieb & Oudeyer, 2018; Gottlieb et al., 2013).

We thus suggested that NERD – a denoising diffusion model trained with RL algorithms – could provide a powerful framework for capturing the brain's process of learning about its own noise in service of minimizing uncertainty. We then trained NERD on an existing DecNef dataset, and projected it into HO space (rather than voxel noise space) using dimensionality reduction techniques that are often used to link neural patterns with representations (mental structures) (Bishop, 2006; Pearson, 1901; Schneider et al., 2023; Steinmetz et al., 2021; Stringer et al., 2019; Vázquez-García et al., 2024). With this approach, we found that the lower-dimensional prior-like uncertainty HORs discovered by NERD could indeed capture individual variation in humans' ability to solve the DecNef task. We believe that the noise distributions learned by NERD under these conditions could thus provide a window into these uncertainty priors as component inputs to the construction of posterior-like HORs about uncertainty. These findings suggest that the brain not only tracks immediate uncertainty, but also builds expectations about uncertainty patterns which likely guide how we learn and make decisions, supporting the proposal developed here that HORs about uncertainty are constructed through a Bayesian-like process.

Summary and final thoughts

Here we've discussed varieties of higher-order representations (HORs), with specific attention to the component ingredients that serve as inputs to the metacognitive processes constructing HORs. We proposed that HORs about uncertainty may be constructed according to a Bayesian-like process, such that reported metacognitive estimates of uncertainty are posterior-like, reflecting a combination of prior-like and likelihood-like HO estimates of uncertainty. We then discussed how various approaches, from the GLMsingle (Prince et al., 2022) and GSN (Kay et al., 2024) approaches to TAFKAP (van Bergen & Jehee, 2021; van Bergen et al., 2015) and our newly-developed NERD model (Azimi Asrari & Peters, 2025; Azimi Azrari & Peters, 2024) may provide exciting paths forward towards studying these component ingredients of uncertainty HORs.

We believe that formulating the construction of uncertainty HORs as Bayesian-like can add crucial clarity to our understanding of the variety of uncertainty-based HORs computed by the brain, as well as their neural correlates and generating computations. We also hope this discussion may also inspire and enable a systematic exploration of HORs in general, beyond those which are about FOR uncertainty, providing crucial insight into the complex relationships between neural patterns and the kinds of mental states they represent.

Conflict of Interest Statement

M.A.K.P. is a consultant for the for-profit entity Conscium, Inc., which seeks to pioneer safe, efficient artificial intelligence and which played no role in this project's conceptualization, analyses, interpretation, or writing. The authors declare no conflicts of interest.

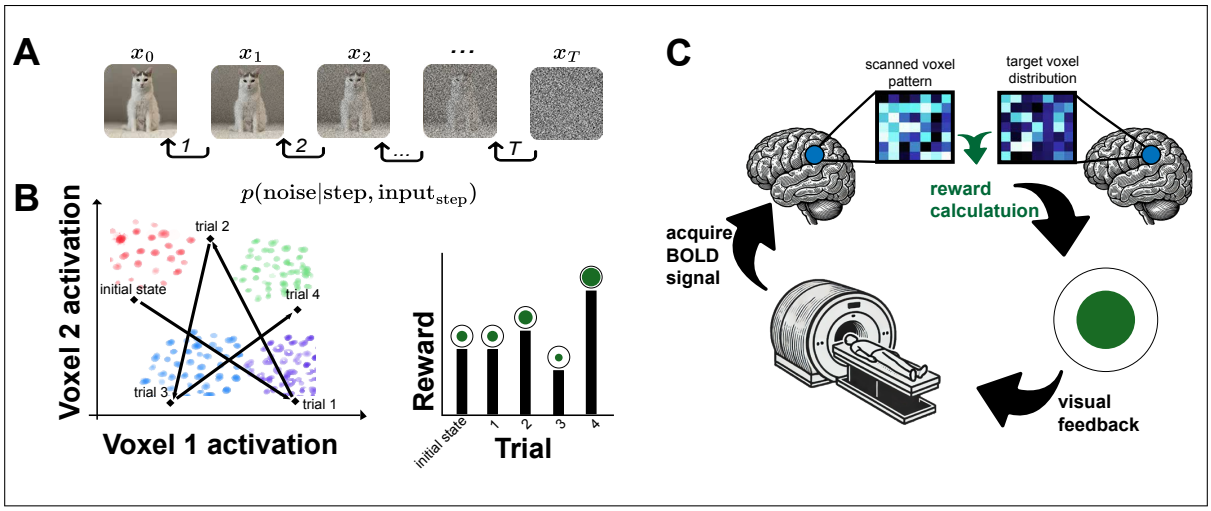

Figure 2: **Cartoons showing the denoising process learned by diffusion models and the closed-loop real-time neurofeedback training procedure.** (A) Denoising diffusion models are trained to learn distributions of pixel noise, conditioned on the denoising step and input image x_T, i.e. $p(noise|step, input_{step})$, in order to denoise the input image such that a new image x_0 from the target distribution can be produced. (B) We have recently hypothesized (Azimi Asrari & Peters, 2025) that this denoising process is undertaken by brains in order to navigate through possible neural patterns in search of a refined goal state, which is likely accomplished through reinforcement learning (RL) in environments where the goal state is not known to the observer. Denoising – or uncertainty reduction – provides a natural candidate mechanism that the brain is equipped to attempt even when the specific goal state is totally unknown. In decoded neurofeedback (DecNef), the observer seeks states which minimize the difference between the current state and the *distribution* of target states, and the degree of match is displayed to the observer as a visualization of computed reward. (C) The closed-loop DecNef procedure involves human subjects learning to denoise their own brain states through RL. Neural response patterns (blood oxygen level dependent [BOLD] signal) are acquired in a given region of interest using functional magnetic resonance imaging (fMRI), compared to a target neural pattern (defined by previous activity patterns), and the degree of similarity between current and target neural state is displayed back to the human participant in the form of a visual feedback circle.

Acknowledgments

This project was supported in part by a fellowship (to M.A.K.P.) from the Canadian Institute for Advanced Research Program in Brain, Mind, & Consciousness and a grant from the Templeton World Charity Foundation ("An adversarial collaboration to empirically evaluate higher-order theories of consciousness", to M.A.K.P.). The funding sources had no role in the design, implementation, or interpretation of the work presented here.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Azimi Asrari, H., & Peters, M. A. K. (2025). Revealing higher-order neural representations of uncertainty with the noise estimation through reinforcement-based diffusion (nerd) model. *arXiv preprint arXiv:2503.14333*. https://arxiv.org/abs/2503.14333

Azimi Azrari, H., & Peters, M. A. (2024). Diffusion models and reinforcement learning: Novel pathways to modeling decoded fmri neurofeedback. *Proceedings of the Cognitive Computational Neuroscience Meeting*.

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in cognitive sciences*, *26*(11), 942–958.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108. https://doi.org/10.1038/s41586-020-2350-5

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Brouwer, G. J., & Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of Neurophysiology*, *106*(5), 2108–2119. https://doi.org/10.1152/jn.00540.2011

Brown, R. (2015). The horor theory of phenomenal consciousness. *Philos. Stud.*, *172*(7), 1783–1794.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Chang, L., Egger, B., Vetter, T., & Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, *31*(14), 2940–2949.e4. https://doi.org/10.1016/j.cub.2021.05.064

Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Front. Psychol.*, *2*, 86.

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2019). Learning to be conscious. *Trends in Cognitive Sciences*, *23*(12), 921–934. https://doi.org/10.1016/j.tics.2019.09.011

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Conscious access to first-order and higher-order representations. *Trends Cogn. Sci.*, *11*(11), 465–472.

Cortese, A., Tanaka, S. C., Amano, K., Koizumi, A., Lau, H., Sasaki, Y., Shibata, K., Taschereau-Dumouchel, V., Watanabe, T., & Kawato, M. (2021). The decnef collection, fmri data from closed-loop decoded neurofeedback experiments. *Scientific Data*, *8*(1), 69. https://doi.org/10.1038/S41597-021-00845-7

De Ridder, D., Vanneste, S., & Freeman, W. (2014). The bayesian brain: Phantom percepts resolve sensory uncertainty. *Journal of Neuroscience*, *34*(46), 15094–15100. https://doi.org/10.1523/JNEUROSCI.3360-14.2014

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown; Company.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE Publications. https://books.google.com/books?id=eVUXBAAAQBAJ

Elliott-Graves, A. (2020). What is a target system? *Biology & Philosophy*, *35*(28), 1–22. https://doi.org/10.1007/s10539-020-09745-3

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*, 1165622.

Favela, L. H., & Machery, E. (2025). Contextualizing, eliminating, or glossing: What to do with unclear scientific concepts like representation. *Mind & Language*.

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of consciousness*, *2020*(1), niz020.

Fleming, S. M. (2023). Metacognitive psychophysics in humans, animals, and ai: A research agenda for mapping introspective systems. *J. Conscious. Stud.*, *30*(9-10), 113–128.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychol. Rev.*, *124*(1), 91–114.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Front. Hum. Neurosci.*, *8*, 443.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.*, *11*(2), 127–138.

Friston, K., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Cognitive Neuroscience*, *8*(2), 144–153. https://doi.org/10.1080/17588928.2016.1267040

Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). Response-based outcome predictions and confidence regulate feedback processing and learning. *eLife*, *10*, e62825. https://doi.org/10.7554/eLife.62825

Gershman, S. J. (2019). The generative adversarial brain. *Front. Artif. Intell.*, *2*, 18.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Goris, R. L., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865. https://doi.org/10.1038/nn.3711

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770. https://doi.org/10.1038/s41583-018-0078-0

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Guggenmos, M. (2022). Reverse engineering of metacognition. *eLife*, *11*, e75420. https://doi.org/10.7554/eLife.75420

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*, e13388.

Hainguerlot, M., Vergnaud, J.-C., & de Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 5602. https://doi.org/10.1038/s41598-018-23936-9

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691. https://doi.org/10.1038/nn1445

Jehee, J. F. M., Ling, S., Swisher, J. D., van Bergen, R. S., & Tong, F. (2012). Perceptual learning selectively refines orientation representations in early visual cortex. *Journal of Neuroscience*, *32*(47), 16747–16753. https://doi.org/10.1523/JNEUROSCI.6112-11.2012

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kammerer, F., & Frankish, K. (2023). What forms could introspective systems take? a research programme. *J. Conscious. Stud.*, *30*(9-10), 13–48.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kay, K. N., Prince, J. S., Gebhart, T., Tuckute, G., Zhou, J., Naselaris, T., & Schutt, H. (2024). Disentangling signal and noise in neural responses through generative modeling. *bioRxiv*. https://doi.org/10.1101/2024.04.22.590510

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as bayesian inference*. Cambridge University Press. https://www.cambridge.org/core/books/perception-as-bayesian-inference/0442F577F5E4CD874FA6819978574C8F

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen.*, *126*(2), 349–370.

LaConte, S. M. (2011). Decoding fmri brain states in real-time. *Neuroimage*, *56*(2), 440–454.

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring.

Lau, H., & Rosenthal, D. M. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.*, *15*(8), 365–373.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, *9*(11), 1432–1438.

Ma, W. J., & Pouget, A. (2009). Population codes, correlations, and coding. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th, pp. 135–144). MIT Press.

Machery, E. (2025). Neural representations: A normative account. *Mind & Language*. https://doi.org/10.1111/mila.12531

Mamassian, P. (2018). Confidence forced-choice and other metaperceptual tasks. *Perception*, *47*(10-11), 1023–1035. https://doi.org/10.1177/0301006618790116

Mamassian, P. (2024). Cassandre: A framework for confidence-aware signal detection in noisy environments [Unpublished manuscript]. *PsyArXiv*.

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol. Rev.*, *129*(5), 976–998.

Mamassian, P., & de Gardelle, V. (2024). The confidence-noise confidence-boost (cncb) model of confidence rating data. *bioRxiv*. https://doi.org/10.1101/2024.09.04.611165

Maniscalco, B., Castaneda, O. G., Odegaard, B., Morales, J., Rajananda, S., Denison, R., & Peters, M. A. K. (2024). The relative psychometric function: A general analysis framework for relating psychological processes. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/5qrjn

Maniscalco, B., Charles, L., & Peters, M. A. K. (2024). Optimal metacognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02510-7

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance sdt model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

Maniscalco, B., & Lau, H. (2016a). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*. https://doi.org/10.1093/nc/niw002

Maniscalco, B., & Lau, H. (2016b). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, *2016*(1), niw002. https://doi.org/10.1093/nc/niw002

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868. https://doi.org/10.1073/pnas.1615773114

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS Computational Biology*, *11*(6), e1004305. https://doi.org/10.1371/journal.pcbi.1004305

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as a metacognitive source of learning speed. *Nat. Rev. Neurosci.*, *16*(12), 721–729.

Michel, M. (2024). The Perceptual Reality Monitoring Theory (1st edition). In M. Herzog, A. Schurger, & A. Doerig (Eds.), *Scientific Theories of Consciousness: The Grand Tour*. Cambridge University Press.

Michel, M., & Lau, H. (2021). Higher-order theories do just fine. *Cognitive Neuroscience*, *12*(2), 77–78. https://doi.org/10.1080/17588928.2020.1839402

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.*, *24*, 167–202.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.*, *84*(3), 231–259.

Odegaard, B., & Shams, L. (2016). The brain's tendency to bind audiovisual signals is stable but not general. *Psychological Science*, *27*(4), 583–591. https://doi.org/10.1177/0956797616628860

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in visual, auditory, and audiovisual perception of space. *PLOS Computational Biology*, *11*(12), e1004649. https://doi.org/10.1371/journal.pcbi.1004649

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Peels, R. (2016). The empirical case against introspection. *Philos. Stud.*, *173*(9), 2461–2485.

Peters, M. A. K. (2020). Confidence in decision-making. *Oxford Research Encyclopedia of Neuroscience*. https://doi.org/10.1093/acrefore/9780190264086.013.371

Peters, M. A. K. (2022). Towards characterizing the canonical computations generating phenomenal experience. *Neurosci. Biobehav. Rev.*, *142*, 104903.

Peters, M. A. K., Balzer, J., & Shams, L. (2015). Smaller = denser, and the brain knows it: Natural statistics of object density shape weight expectations. *PLOS ONE*, *10*(3), e0119794. https://doi.org/10.1371/journal.pone.0119794

Peters, M. A. (2025). Introspective psychophysics for the study of subjective experience. *Cerebral Cortex*, *35*(1), 49–57.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., et al. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature human behaviour*, *1*(7), 0139.

Pospisil, D. A., & Pillow, J. W. (2024). Revisiting the high-dimensional geometry of population responses in visual cortex. *bioRxiv*. https://doi.org/10.1101/2024.02.16.580726

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, *11*, e77599.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, *159*, 271–295. https://doi.org/10.1007/s11229-007-9208-3

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist*, *76*(9), 1445–1453. https://doi.org/10.1037/amp0000852

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., ... Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Psychonomic Bulletin & Review*, *29*(5), 1553–1562. https://doi.org/10.1177/17456916221075615

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169–192.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press. https://www.amazon.com/Consciousness-Mind-David-Rosenthal/dp/0198236964

Rosenthal, D. M. (2012). Higher-order awareness, misrepresentation and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1424–1438. https://doi.org/10.1098/rstb.2011.0353

Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, *617*, 360–368. https://doi.org/10.1038/s41586-023-06031-6

Schneider, S. (2020). Mental representation (E. N. Zalta, Ed.). https://plato.stanford.edu/entries/mental-representation/

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philos. Rev.*, *117*(2), 245–273.

Schwitzgebel, E. (2011). *Perplexities of consciousness*. MIT Press.

Serences, J. T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L. T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, *44*(1), 223–231. https://doi.org/10.1016/j.neuroimage.2008.07.043

Series, P. M., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 668. https://doi.org/10.3389/fnhum.2013.00668

Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? a comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General*, *153*(3), 656.

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fmri neurofeedback without stimulus presentation. *Science*, *334*(6061), 1413–1415. https://doi.org/10.1126/SCIENCE.1212003

Smith, M. A., & Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *Journal of Neuroscience*, *28*(48), 12591–12603. https://doi.org/10.1523/JNEUROSCI.2929-08.2008

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*(5026), 1380–1386.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., ... Harris, T. D. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588. https://doi.org/10.1126/science.abf4588

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. https://doi.org/10.1038/nn1669

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365. https://doi.org/10.1038/s41586-019-1346-5

Tarr, M. J., & Vuong, Q. C. (2002). Visual object recognition. *Steven's handbook of experimental psychology*, *1*, 287–314.

Thornton, M. A., & Tamir, D. I. (2024). Neural representations of situations and mental states are composed of sums of representations of the actions they afford. *Nature Communications*, *15*, 620. https://doi.org/10.1038/s41467-024-00620-0

van Bergen, R. S., & Jehee, J. F. M. (2021). Tafkap: An improved method for probabilistic decoding of cortical activity. *bioRxiv*. https://doi.org/10.1101/2021.03.04.433946

van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*, 1728–1730. https://doi.org/10.1038/nn.4150

Vázquez-García, C., Martínez-Murcia, F., Segovia Román, F., & Górriz, J. M. (2024). A review of latent representation models in neuroimaging. *arXiv preprint arXiv:2412.19844*. https://arxiv.org/abs/2412.19844

Vilarroya, O. (2017). Neural representation: A survey-based analysis of the notion. *Frontiers in Psychology*, *8*, 1458. https://doi.org/10.3389/fpsyg.2017.01458

Von Eckardt, B. (2012). The representational theory of mind. *The Cambridge handbook of cognitive science*, *1*(29-50).

Walker, E. Y., Pohl, S., Denison, R. N., Barack, D. L., Lee, J., Block, N., Ma, W. J., & Meyniel, F. (2023). Studying the neural representations of uncertainty. *Nat. Neurosci.*, *26*(11), 1857–1867. https://doi.org/10.1038/s41593-023-01444-y

Watanabe, T., Sasaki, Y., Shibata, K., & Kawato, M. (2017). Advances in fmri real-time neurofeedback. *Trends in cognitive sciences*, *21*(12), 997–1010.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Williams, A. H., & Linderman, S. W. (2021). Statistical neuroscience in the single trial limit. *Current Opinion in Neurobiology*, *70*, 193–205. https://doi.org/10.1016/j.conb.2021.10.008

Winter, C. J., & Peters, M. A. (2022). Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery. *Attention, Perception, & Psychophysics*, *84*(1), 161–178.